International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE Open Access

Students' complex trajectories: exploring degree change and time to degree

João Pedro Pêgo[1,2], Vera Lucia Miguéis[3*] and Alfredo Soeiro[4]

*Correspondence:
vera.migueis@fe.up.pt

[1] Faculdade de Engenharia, da Universidade do Porto, Departamento de Engenharia Civil, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[2] CIIMAR, Interdisciplinary Centre of Marine and Environmental Research, Terminal de Cruzeiros de Leixões, Av. General Norton de Matos s/n, 4450-208 Matosinhos, Portugal
[3] Faculdade de Engenharia da Universidade do Porto, INESC TEC, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[4] Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

Abstract

The complex trajectories of higher education students are deviations from the regular path due to delays in completing a degree, dropping out, taking breaks, or changing programmes. In this study, we investigated degree changing as a cause of complex student trajectories. We characterised cohorts of students who graduated with a complex trajectory and identified the characteristics that influenced the time to graduation. To support this predictive task, we employed machine learning techniques such as neural networks, support vector machines, and random forests. In addition, we used interpretable techniques such as decision trees to derive managerial insights that could prove useful to decision-makers. We validated the proposed methodology taking the University of Porto (Portugal) as case study. The results show that the time to degree (TTD) of students with and without complex trajectories was different. Moreover, the proposed models effectively predicted TTD, outperforming two benchmark models. The random forest model proved to be the best predictor. Finally, this study shows that the factors that best predict TTD are the median TTD and the admission regime of the programme of destination of transfer students, followed by the admission average of the previous programme. By identifying students who take longer to complete their studies, targeted interventions such as counselling and tutoring can be promoted, potentially improving completion rates and educational outcomes without having to use as many resources.

Keywords: Complex trajectories, Machine learning model, Time to degree, Programme transfer

Introduction

Higher education enrolment has increased in developing countries in recent decades (Barakat & Shields, 2019). In these countries, the democratisation of access to higher education has been promoted, eliminating the elite exclusivity of the university. This has happened because of the need to increase the number of highly qualified human resources to guarantee economic competitiveness and other factors related to the alteration of individual life. Indeed, investment in education has become increasingly important for young people as they realise that educational qualifications can improve their chances of finding better job opportunities (Wang, 2021). In parallel, a political

movement to increase the number of young people entering higher education has been observed (Dias, 2015).

In several countries, namely Portugal, this pressure for expansion has led to increased system diversification, resulting in a binary system comprising university and polytechnic studies (Sousa, 2021), as well as in the emergence of the private sector (Teixeira et al., 2022). This process of massification has also led to the constitution of a large number of different programmes (Dias, 2015). In addition, the attempt to get more students into higher education has led to the provision of financial support for less advantaged students to cover tuition fees and other education expenses (Biffl & Isaac, 2002).

The increasing diversity of the student population and of higher education institutions and programmes has encouraged the diversification of trajectories (Haas & Hadjar, 2020). Indeed, the process of higher education is not always linear (Rosenberg et al., 2018), and students have been experiencing increasingly complex trajectories, including dropping out, stopping out for a time, transferring between programmes or institutions, enrolling part-time, and taking longer to conclude degree programmes (Goma, 2023; Rosenberg et al., 2018).

These complex processes result from various factors within and across multiple dimensions of the educational system. Each student's psychological, social, and cognitive features may impact their progress. In addition, the background and the level of preparation of students also constitute important determinants of their trajectories. It is also widely recognised that students' experiences, university environment, and professors may play a critical role in their trajectories (Bowman & Holmes, 2018; Tinto, 1994; Xerri et al., 2018; Xie et al., 2015).

A significant part of students' complex trajectories is transferring to other programmes (Hovdhaugen, 2009). This is a particular issue in countries such as Portugal, where access to higher education is guided by a *numerus clausus* system established through a national public tender to prioritise the admission of students with higher access classification (Ferrão & Almeida, 2018). This leads to more than half of Portuguese students not accessing the course/establishment pair assumed as their first choice when applying for higher education (Casanova & Almeida, 2016). Thus, many students who end up being placed in the other options find ways to change course. Some resort to the course transfer system, which has vacancies for course changes every curricular year. Other students choose to re-apply the next academic year, hoping to gain access to their course of choice (Almeida et al., 2016). Similar scenarios happen in different countries, with students using the first year as a bridge to their preferred course (Okun et al., 2009).

It is important to note that changing programmes involves two decisions. The first decision is to leave the original programme. The second decision is to either leave higher education or change to an alternative degree programme within higher education (Tieben, 2020). The determinants of course transfer have been shown to differ from those of dropping out (Ferrão & Almeida, 2018). The conditions of the institution play a dominant role in the decision to quit a course (Berger & Braxton, 1998), while the opportunities and conditions outside the university are relevant considerations when deciding to drop out (Tieben, 2020).

In this paper, we focus our analysis on students with a complex trajectory, namely those who have transferred from the programme in which they enrolled at university

to a different one. We aim to propose a method that enables programme managers and counselling and tutoring service providers to determine the time to degree (TTD) for each student. Studies show that transfer students are more likely to either not complete their degrees or take longer to complete their degrees when compared to students who have not experienced a complex trajectory (Townsend & Wilson, 2009). Transfer students are forced to navigate through complicated systems in order to take advantage of course credits and to enrol in courses, which can lead some of them to feel unwelcome or even marginalised by their new institution or degree programme (Chin-Newman & Shaw, 2013). Transfer students can also struggle with social integration as they try to find their place in a new context (Utter & DeAngelo, 2015). These barriers to social and academic integration can lead to transfer students taking longer to conclude their studies.

Having identified the students most likely to have a longer TTD, higher education institutions can design policies to prevent longer trajectories. For example, institutions may ask students who are predicted to take shorter routes to interact and share experiences with the other students, as a way to shorten the expected routes of the latter.

Through this paper, we seek to provide many valuable contributions to the literature. First, we aim to explore students' trajectories, a topic still very much unexplored in the literature, as studies focusing on individual students and their specific trajectories are rare. Moreover, we seek to study complex trajectories, particularly those involving a transfer, a less prominent topic that has been mostly neglected. In addition, we aim to contribute to the literature by exploring the TTD of students after they transfer to another programme, which, to the best of our knowledge, has also not been addressed before. We seek to estimate the time that newly enrolled transfer students need to conclude the new programme by applying several machine learning techniques, namely random forests, bagged trees, and boosted trees. The application of ensemble methods, such as those previously mentioned, to the educational data mining field is still in the early stages, although their predictive performance is generally high. Lastly, we aim to provide decision-makers with information on the factors that impact the TTD of students by evaluating the importance of student-related variables to the prediction model.

This paper is structured as follows. The following section presents related studies in order to emphasise the contributions of the current study. "Methodology" section introduces the methodology and data used in the current study, the variables included in the proposed model, and the criteria used to evaluate the performance of the model. "Results" section presents the results, which are discussed in "Discussion" section. "Study limitations" section highlights the limitations of the study and "Conclusions and future work" presents the study's conclusions and ideas for future research.

Literature review

Student trajectory in higher education refers to the "progression through higher education including all transitions (e.g. from undergraduate to graduate studies) and states (e.g. enrolment patterns such as part-time vs full-time enrolment) within a certain period (e.g. academic year or 3-year life period)" (Haas & Hadjar, 2020). Giani (2015) divided the trajectory of higher education students into seven stages: application, acceptance,

enrolment, persistence/transfer, attainment, graduate school entry, and graduate school attainment.

The literature has incipiently addressed the study of student trajectories in higher education (Haas & Hadjar, 2020). A literature review on student trajectories between 1999 and 2018 identified only 27 articles (Haas & Hadjar, (2020). Most of these studies address the reality of USA institutions and use nationally representative large-scale data. The availability of data, particularly longitudinal student data, may be one reason for this lack of studies on student trajectories (Haas & Hadjar, 2020). This type of study deserves more attention, particularly because of the developments in higher education in the last decades. Facilitated access to higher education for underrepresented social groups (Hadjar & Becker, 2009) has encouraged the expansion, diversification (Schofer & Meyer, 2005), and heterogeneity of student populations, which should motivate studies of student trajectories.

The literature on student trajectories has adopted three distinct research designs (Haas & Hadjar, 2020). The first set of studies has focused on describing trajectory types and patterns (Robinson, 2004). The second and third sets of studies have focused on answering specific questions concerning students' attributes and determining who follows which trajectories and why (Giani, 2015; Goldrick-Rab, 2006).

The study of complex trajectories is a niche within the topic of student trajectories. Complex trajectories include dropping out, stopping out for some time, transferring between programmes or institutions, enrolling part-time, and taking longer to conclude a degree programme (Goma, 2023; Rosenberg et al., 2018). Despite the relevance of longitudinal studies on complex trajectories, the literature has only focused on very specific issues such as dropping out, neglecting the fact that students may transfer between study programmes or institutions, interrupt their studies, or slow down the pace of study (Haas & Hadjar, 2020). Indeed, dropping out has been the most explored dimension (e.g. Berzenski, 2021; Tieben, 2020). Most of the studies on dropping out have focused on predicting whether a student is prone to quit their studies. In contrast, few studies have sought to explore the transfer of students between programmes (e.g. Rodríguez-Gómez et al., 2016). However, empirical studies have shown that the determinants of programme transfer differ from those of dropping out (Ferrão & Almeida, 2019). For example, Terenzini et al. (1981) and Yi (2008) showed the distinct impact of student characteristics on events such as transferring and dropping out. In this context, we may conclude that there is a need for research into the complex trajectories of higher education students, such as those who transfer between programmes.

In parallel, the literature on TTD, i.e. the number of years it takes for a student to complete a higher education degree, is also scarce, particularly in terms of its estimation (Bhaskaran et al., 2017). TTD is a relevant metric because it can show the efficiency of an educational system (Rayner & Papakonstantinou, 2022). The longer it takes for a student to graduate, the more resources are used by higher education institutions and by the student to achieve their final goal (Iatrellis et al., 2020). In the USA, about 41% of higher education students fail to graduate within six years (Basavaraj & Garibay, 2019). In Europe, only 23% to 30% of higher education students graduate within the expected time (Boegeholz et al., 2022). In this context, early estimation

of each student's TTD may be paramount. With such an estimate, institutions may design customised actions that prevent long trajectories. This is particularly relevant for students who have already experienced a programme transfer, as they already took longer to reach their preferred course.

For example, Hailikari et al. (2019) used interviews to categorise first-year students into six profiles and concluded that there were significant differences in graduation times among these profiles. They also found large differences in the completion rates of master's degrees between professional and non-professional fields, with students from the humanities tending to prolong their studies due to a fear of unemployment. Rayner and Papakonstantinou (2022) also explored the TTD of students. In particular, this study sought to identify the predictors of TTD for undergraduates and concluded that the most relevant are the gender, the admission rank, the number of discipline majors, and the level of academic achievement.

Concerning the methodology adopted by the studies exploring educational data, recently machine learning methods have been gaining momentum (Karalar et al., 2021). Aldowah et al. (2019) reviewed 402 articles and identified the machine learning techniques used in educational data mining and learning analytics. According to this literature review, most of the studies adopt classification techniques (26.25%) and clustering (21.25%) (Aldowah et al., 2019). Romero and Ventura (2020) listed the following machine learning approaches, among others: causal mining to relate student behaviour to learning, academic failure, or dropping out; clustering to group materials or students; prediction of student performance and student behaviour; and social network analysis to interpret the structure and relationships in collaborative activities. Sghir et al. (2022) highlighted that predicting student performance dominates the field, followed by identifying at-risk students.

Regarding the research on predictive analytics in higher education, Sghir et al. (2022) reviewed several papers published in the last decade (2012–2022) in order to identify the algorithms and goodness-of-fit measures most commonly used. They concluded that artificial neural networks attained the best performance in classification problems, followed by random forests (Boehmke & Greenwell, 2019) and gradient boosting (Friedman, 2001, 2002; Mason et al., 1999). Decision trees, naive Bayes, ensemble methods, and k-nearest neighbours were also identified as popular algorithms. Concerning regression problems, single and multiple linear regression algorithms have been used in prediction tasks. For clustering problems, (Sghir et al., 2022) identified seven articles using the k-means algorithm (MacQueen, 1967). In terms of performance measures, the ones most commonly used for classification tasks were, in descending order, frequency, accuracy, F-measure, recall, precision, area under the ROC curve (AUC), kappa, sensitivity, specificity, and Mathew Correlation Coefficient (MCC). For regression problems, the measures used include Pearson's R, the root mean square error (RMSE), the predictive mean square error (pMSE), and the predictive mean absolute percentage correction (pMAPC).

Concerning the variables used in predictive learning analytics, several are commonly associated with predicting graduation, dropping out, and academic performance. Sghir et al. (2022) classified predictor variables into five classes, as follows. Prior academic data includes the student's records in secondary education (Berzenski, 2019; Tieben, 2019)

or admission information such as admission exam grades (Carreira & Lopes, 2019) and scientific area (Carreira & Lopes, 2019; Rodríguez-Gómez et al., 2016). Demographic characteristics include personal data such as gender (Hashim et al., 2020; Martins et al., 2019; Sánchez-Gelabert et al., 2020), ethnicity (Berzenski, 2019; Monaghan, 2019), and age (Hashim et al., 2020; Monaghan, 2019; Tumen et al., 2008). They also include the socio-economic context of the student (e.g. the country of origin/residence) (Carreira & Lopes, 2019; Rodríguez-Gómez et al., 2016) and the educational and occupational level of the student's family (Sánchez-Gelabert et al., 2020; Hashim et al., 2020). Academic data pertains to the student's performance at the higher education institution. Other commonly employed predictors (Brezavšcek et al., 2017) include the area of study (Hashim et al., 2020), the number of completed credit points (Berzenski, 2019; Martins et al., 2018), the grade point average (GPA) Berzenski (2019); Hashim et al. (2020), and the time spent at each programme stage. Behavioural features are mostly employed with data retrieved from learning management systems Aldowah et al. (2019); Prenkaj et al. (2021); Sghir et al. (2022), as they are easily accessible (Romero & Ventura, 2020). The motivation of the student is a relevant factor for success (Pardo et al., 2017; Wong & Chiu, 2019) and is a good example for the last category, which is psychological data.

In the context of higher education trajectories, (Haas & Hadjar, 2020) defined three levels of trajectory predictors. The macro level includes factors deriving from the regional and national structures of the higher education system (e.g. fees and financial aid). The meso level considers factors related to the organisational structures of higher education institutions (e.g. offering mentorship and the size of the programme). The micro-level includes factors that depend on the student's context, such as socio-demographics, expectations, and academic preparation.

Methodology

Research questions

In this study, we aim to examine the complex trajectories of higher education students, namely those that involve a change of programme. For these trajectories, we intend to estimate the TTD after the change to a different programme. We propose to use machine learning models to obtain accurate estimates of the TTD of students and help identify the factors that are most relevant to distinguish students presenting different progressions.

The proposed solution should benefit higher education institutions and students by allowing an early and accurate prediction of TTD and the subsequent implementation of remedial plans to reverse scenarios where a long TTD is expected. The identification of students with a lower TTD should also contribute to this inversion, for example, by encouraging more interaction between these students and those with a potentially higher TTD. Sharing experiences and practices may help to mitigate the difficulties that some students may encounter. Overall, this may enable institutions to maintain their reputation for academic excellence.

More specifically, we formulated the following research questions:

- RQ1: Is the TTD of students who transfer to another programme different from that of students who have a non-complex trajectory?

- RQ2: Can a machine-learning model that integrates variables characterising students' previous trajectories infer the TTD after a transfer?
- RQ3: What are the most important factors when predicting the TTD of students with complex trajectories?

Research design

The methodology proposed in this study is illustrated in Fig. 1.

This study focuses on the students who completed at least a degree programme during the analysis period. Moreover, it only considers the academic trajectory until the first degree completion. Having set the sample of analysis, the first step of the proposed methodology is to distinguish the students who transferred from one programme to another from those who did not transfer.

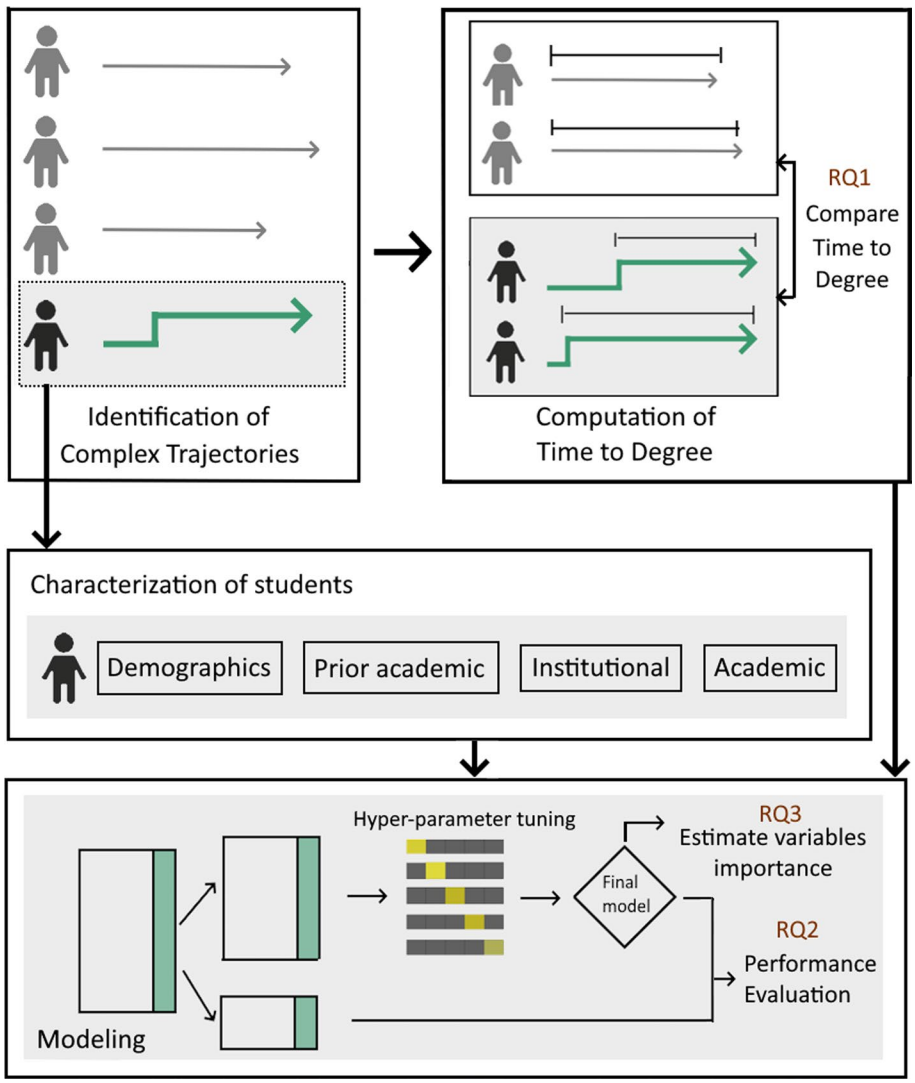


Fig. 1 Overview of the methodology

Secondly, the TTD is computed for each student, i.e. the difference between the time of conclusion and the time of enrolment in the final programme. It should be noted that in the present study, the TTD is expressed in years, since only yearly data was available. It is also worth pointing out that, as noted in the literature (see "Literature review" section), barriers to social and academic integration may mean that transfer students take longer to complete their studies. On the other hand, transfer students may be able to use credits from previous programmes, which may also affect the length of their studies. This computation is done for the groups of students who transferred from one programme to another and for those who did not transfer. Having collected this data, we propose using a Z-test to answer the first research question and thus establish the statistical relevance of the difference in the TTD of the two groups.

Next, we propose characterising students after they have transferred to a new programme. Following the categorisation proposed by Sghir et al. (2022), we suggest including variables related to the prior academic background and demographics. In addition, we propose to describe students based on their academic-related variables (Sghir et al., 2022). We also propose to characterise both the original and the new programme, i.e. institution-related variables (Sghir et al., 2022). Table 4 lists all the variables used to define a student after a change. Figure 2 and the third column of Table 4 help to better understand the time frame corresponding to each variable. Some variables concern the period before enrolling in the university (h_s), namely the type of high school and the grades on the national admission exams. Other variables refer to the period between the admission to the university and the moment of the programme change (Y_p), e.g. the number of enrolments, the number of programmes enrolled in that period, or the percentage of time the student was working or displaced. Some other variables refer to the most recent academic year before the change (Y_r), e.g. the programme of origin or the cumulative number of credits completed before the change. Finally, we also accommodate variables collected at the moment of the transfer (i_c), e.g. final programme faculty, final programme duration, and whether the student has a scholarship or was working at the moment of the change.

It is worth noting that, because machine learning algorithms do not generally accept data with missing values, we propose to exclude from the study students for which the data is not complete. Thus, students presenting one or more missing values were not considered for the training of the predictive models. Moreover, categorical features were one-hot encoded, while ordinal features were ordinally encoded.

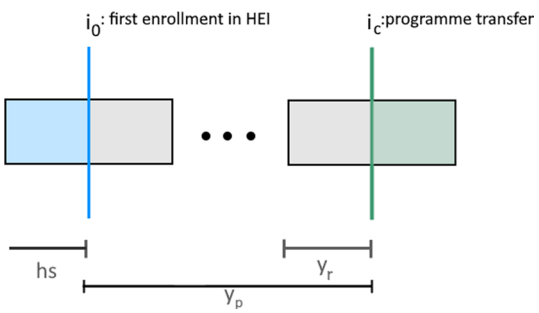


Fig. 2 Overview of the methodology

In the last stage, we propose using a data mining prediction model that uses the variables introduced in Table 4 as independent variables and the TTD as the dependent variable. Thus, the second research question is answered. Although the TTD is discrete, we propose to treat this problem as a regression problem. Considering a predictive model based on classification algorithms would limit the scope of applicability of the model, as the model would only be able to predict the target values observed in the training dataset. Using a regression algorithm to train the model overcomes this limitation. In addition, students' average grades exhibit a degree of continuity, although they were rounded to give a final whole number grade.

Following several studies on education (Casuat & Festijo, 2019; Cortez & Silva, 2008), we propose using the decision tree (DT), random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP) regression algorithms. These are among the most popular regression algorithms and can be applied to relatively small datasets.

DTs are easy to apply, effective, and fast to train. The hierarchical tree structure resembles a human decision-making process, making DTs easy to understand (Czajkowski & Kretowski, 2016). They are also a white box algorithm, meaning they are explainable. The RF algorithm is an ensemble algorithm where several DTs are generated from a random vector sampled independently (Breiman, 2001) and combined by averaging the results to produce one single predicted value. RFs have good predictive performance and have some level of interpretability. They are a good choice when the number of predictive attributes is large, as is the case of the present study, though at a high computational cost (Moreira et al., 2018). SVMs aim to find a hyperplane that best fits the data while minimising the margin of error, making them a powerful tool for non-linear regression. One of the strengths of SVMs is the ability to solve small sample, non-linear, and high dimensional pattern recognition problems while being memory efficient (Wang et al., 2016). An MLP, or artificial neural network, attempts to reproduce the functioning of the human brain. Each node in an MLP is equivalent to a neuron in the human brain. MLPs perform well in many real-life problems, even non-linear problems, and are robust to noise. MLPs are hard to interpret due to the lack of mathematical foundation and hidden layers and their training usually comes at a high computational cost (Moreira et al., 2018).

We propose tuning the hyperparameters using an exhaustive grid search with stratified k-fold cross-validation, where $k = 5$. Thus, the gathered data set is split into train and test datasets, the latter representing 20% of the original one. The training dataset is used to optimise the hyperparameters of each model, while the test dataset is used to assess the performance of the models.

Choosing adequate goodness-of-fit metrics is crucial for the performance evaluation of the models. We recommend the use of the RMSE, the mean absolute error (MAE), the coefficient of determination (R^2), and the mean absolute percentage error (MAPE). We suggest computing the confidence intervals at a confidence level of 95% by bootstrapping the test set predictions for the regression metrics. Regression models should be tuned to guarantee the lowest RMSE.

Finally, we propose following the permutation feature importance approach to identify the importance of the features, answering the third research question. This model inspection technique is recommended for opaque models such as RFs (Breiman, 2001).

It measures the impact of each feature on the model's performance by randomly permuting the values of a single feature and observing the resulting change in the model's predictive goodness-of-fit metric. The process involves different steps, as follows. First, the model is trained on a dataset with all features intact, and its performance metric (in this case, the RMSE) is recorded as the baseline. Then, the values of one feature are shuffled randomly and the dataset is passed through the trained model again to obtain the new value of the performance metric. The difference between the baseline metric and the permuted metric quantifies the importance of that feature. The more significant the drop in performance (in this case, the greater the decrease in the RMSE) after permuting a feature, the more influential that feature is considered to be. By evaluating the permutation feature importance for all features, it is possible to identify which variables have the highest impact on the model's performance and gain insights into their relative importance.

It should be noted that in order to prevent overfitting, in addition to cross-validation, we chose to select simpler models with fewer parameters, like linear regression or simple DTs. We also used an ensemble algorithm (RF), which is less prone to overfitting. We minimised the risk of overfitting in the case of the neural networks through L1 regularisation and in the case of SVMs through C regularisation.

Case study and data description

This study uses data from the University of Porto (U.Porto), a Portuguese public research university. This university has approximately 34,000 students, 3400 academic staff and researchers, and offers undergraduate (bachelor) and graduate (master and doctorate) programmes in several fields, such as engineering, humanities, law, and medicine.

Students' data was collected from the information system of the University of Porto after passing through a rigorous procedure related to data protection and ethical issues. This data covers the demographics and the prior academic, institutional, and current academic data of each student. Students' data was anonymised to ensure individual students could not be identified.

The data gathered corresponds to yearly student information for those enrolled for the first time at the University of Porto between 2005 and 2015 in an entry-level degree, i.e. a bachelor's (B) or an integrated master's (IM) degree. The data available encompasses all the academic information of 52,822 students, obtained until 2020. The 54 courses offered by the 14 faculties of the University of Porto are represented in the dataset, of which 36 are bachelor's degrees and 18 are integrated master's degrees.

Figure 3 presents the ten most frequent trajectories in the period under review, which cover about 93% of all trajectories. "Not Enrolled" refers to the period in which a student who had previously enrolled at the university was not enrolled in any programme. "Final Programme" refers to the period in which the student was enrolled in the programme where they graduated. The period in which the student was enrolled in a programme other than the final programme is labelled "Other Programme". Finally, the period labelled as "Graduated" refers to the year of graduation and subsequent years. It is worth noting that the figure's time axis goes up to 16 years, as this corresponds to the longest academic trajectory available in the dataset. Among the most frequent trajectories, about two thirds lasted five, three, or six years, corresponding to undergraduate (three

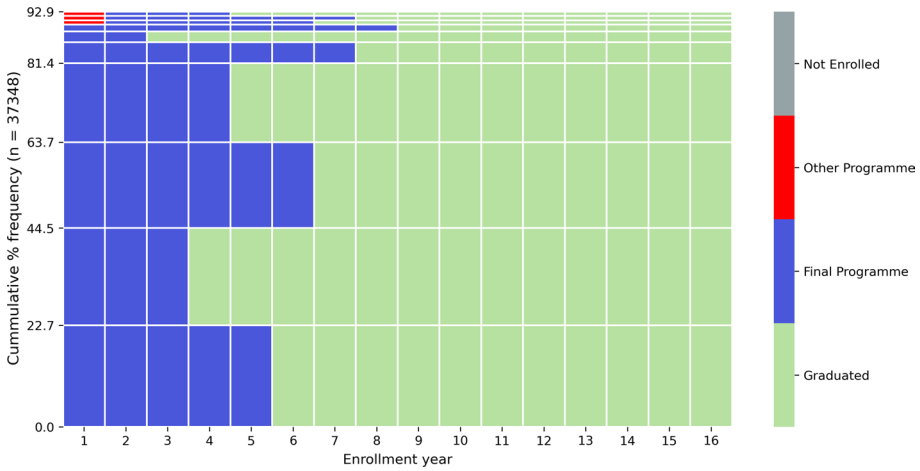


Fig. 3 The 20 most frequent trajectories of students who finished a degree

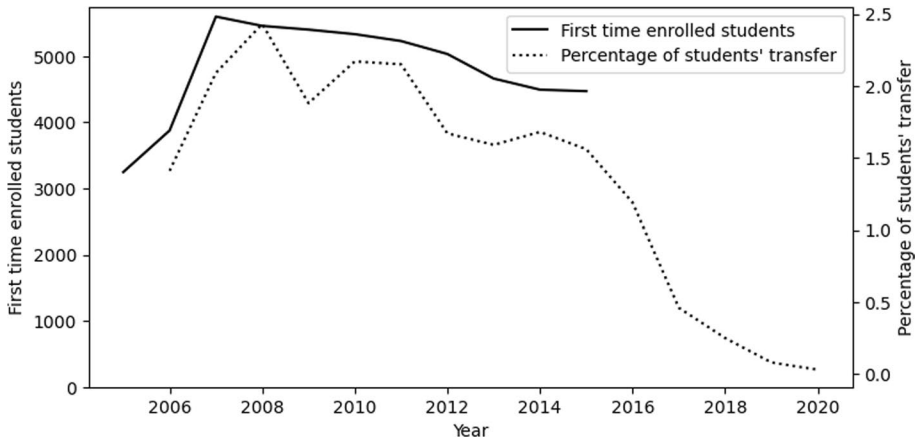


Fig. 4 Number of students enrolled at the University of Porto for the first time and number of students who changed programme

years) and integrated master's (five and six years) students. In addition, there are also frequent trajectories that lasted four years, which corresponds to a bachelor's degree taking one additional year to complete. The remaining popular paths show that about 20% of the students needed extra time to graduate. The complex trajectories, i.e. those with a different programme in the first year, were the least frequent in this top ten.

Looking at the annual enrolments of new students over the analysed period, shown in Fig. 4, it is possible to see that this number increased significantly between 2005 and 2007 and has gradually decreased since then. This can be partly explained by the economic crisis that Portugal experienced between 2010 and 2014. Nevertheless, the number of students enrolled at the University of Porto has always been higher than 3000.

If we focus on students who changed programmes and successfully graduated, Fig. 4 shows that they are a minority. In fact, each year, the maximum percentage of enrolments reflecting a change of programme in relation to the total number of re-enrolments is around 2.43%. Nevertheless, if we focus on all the students who transferred at least once and graduated, this corresponds to 2743 students, or 7.2% of the total number

Table 1 COMPLEX trajectory example

Year	Programme name	Admission regime	Admission average	Credits enrolled	Credits completed	Finished?
1	B in Comm Sci: Journalism,...	RG	16.38	60.0	28.5	No
2	B in Comm Sci: Journalism, ...	RG	16.38	91.5	52.5	No
3	B in Comm Sci: Journalism, ...	RG	16.38	50.0	50.0	No
4	B in Applied Languages	RG	15.74	60.0	54.0	No
5	B in Applied Languages	RG	15.74	66.0	48.0	No
6	B in Applied Languages	RG	15.74	66.0	66.0	Yes

of students who were admitted to the University of Porto between 2005 and 2015 and graduated by 2020. It should be noted that the significant decline in the percentage of student transfers after 2015 is due to the fact that the data collected does not include new students from 2015 onwards, thus reducing the number of potential transfers. In addition, our study focuses only on students who already graduated, so it is possible that more students transferred in the last years of the period analysed but are not reflected in this graph.

With regard to transfers, it is important to look briefly at the programmes of origin and destination. Table 5 in the appendix lists the programmes from which students have transferred to other programmes, listed in descending order of frequency. The top programmes in terms of transfers are engineering and those related to health. The table also shows the three most frequent destination programmes and the respective frequency of students who made this transition. There seems to be a repeating pattern of transfers between programmes, since the most frequent destination courses represent the vast majority of transfers. An extreme example of this is the case of the integrated master's degree in dental medicine, where 95% of the students who transfer are destined for medicine.

Table 1 illustrates an example of a complex trajectory of a student. More specifically, it shows the academic path of a student who enrolled in two programmes at the University of Porto, successfully completing the second. Over the course of six years, the student enrolled in two programmes, completing credits in both. For this student, the initial programme was a bachelor's degree in Communication Sciences: Journalism, Public Relations, Multimedia and the final programme was a bachelor's degree in Applied Languages, which they completed in three years. The moment i_0 corresponds to the beginning of year one, while the moment i_c corresponds to the beginning of year six. Thus, Y_P corresponds to five years. Variables preceding admission to the university are identified in Table 4 as h_s. Variables in the period Y_p include data collected during the time at the university before the transfer. Variables marked with Y_r were collected at the beginning of the year preceding the change, i.e. year five in this particular case.

Results

After identifying the students who transferred between programmes at least once in the analysed period, i.e. the complex trajectories, we computed the TTD of the students who transferred and of those who did not.

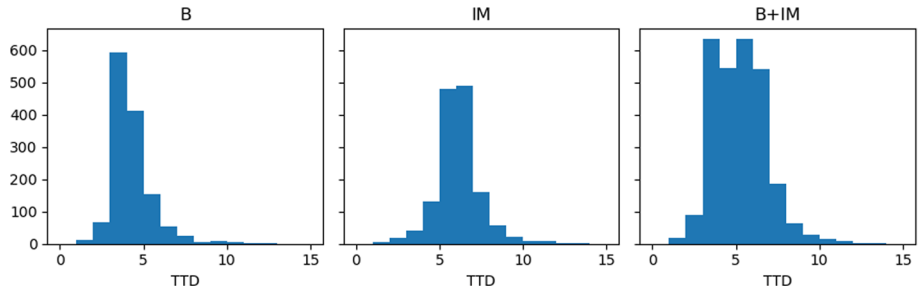


Fig. 5 TTD of students who transferred between programmes per degree type

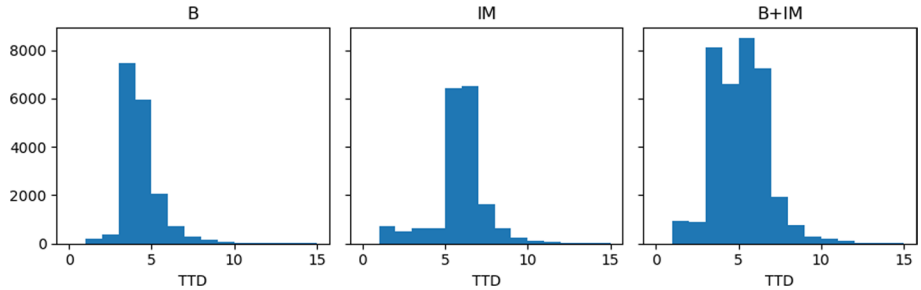


Fig. 6 TTD of students who did not transfer between programmes per degree type

Figures 5 and 6 show the TTD of students who transferred between programmes and those who did not, distributed by degree type. It should be noted that, by definition, an integrated master's degree takes longer than a bachelor's degree, since an integrated master's degree combines a bachelor's and a master's degree. Concerning the bachelor's degrees, the TTDs have a similar distribution across both figures. However, there is a more significant asymmetry to the right in the case of the TTD distribution of the students who changed programmes. In the case of the integrated master's degrees, the distribution for the two populations is more distinct. In this case, it is more evident that the TTD is generally higher in the trajectories with a transfer. The results of the Z-test with a one-sided alternative (see Table 2) show that the null hypothesis, i.e. there is no significant difference between the means of two populations, is rejected for both types of degree, i.e. integrated master's and bachelor's. We can therefore answer the first research question and state that the TTD is different for students who experience a complex trajectory and those who do not. This fact corroborates the literature and emphasises the need to develop specific models to

Table 2 Z-test results

		Average	Variance	Z-value	p-value
Integrated master's	Complex	5.61	1.83	1.64	0.00
	Non-complex	5.43	2.63		
Bachelor's	Complex	3.76	1.60	1.64	0.01
	Non-complex	3.84	1.64		

predict TTD for students who have undergone a course change. While in the case of integrated master's degrees the time taken by a transfer student to complete the programme is longer than that of a non-transfer student, the opposite seems to be true for undergraduate degrees.

To characterise the students who transferred between programmes, we based our analysis on the variables introduced in "Methodology" section. Some descriptive statistics are presented in Table 4. Most of these students are female, do not work, live in their regular house (not displaced), and do not have a scholarship. It is interesting to note that the average age of the students at the time of transfer is 19.47 years, which means that most of them transfer between programmes in their second year of studies, as they are usually 18 years old when admitted to the university. After cleaning the data and excluding observations with missing values, 2047 students remained in the dataset.

Figure 7 shows the TTD values for the test dataset using the programme duration as the benchmark, i.e. the number of years foreseen in the study plan. A second benchmark prediction was made using the programmes' median TTD, computed with the data from the training dataset. The plot for the latter is similar to Fig. 7, as the difference between the two values is small for most programmes. The graph shows that even though the actual values are distributed along a range of TTDs from one to ten years, the values predicted by the benchmark cover a smaller range, between two and a half and six years, with the smaller values corresponding to the bachelor's degrees and the larger values corresponding to the integrated master's degrees, namely the one in medicine. Table 3 shows the performance metrics of both benchmarks, which do not differ much, with a slightly better performance in the programme duration benchmark. The results

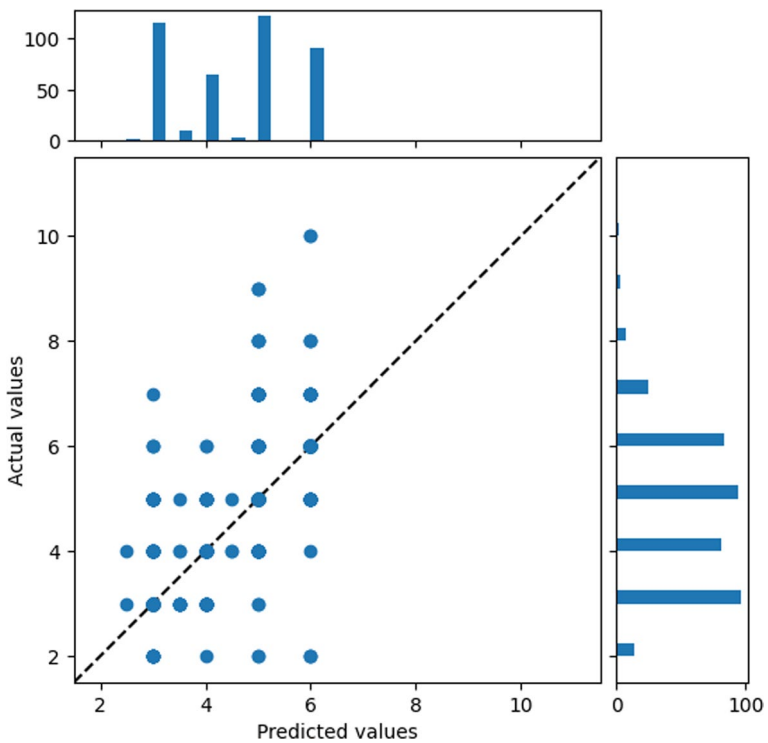


Fig. 7 Benchmark predictions (programme duration)

Table 3 Performance metrics for benchmarks (BM) and models (mean and 95% confidence interval)

Benchmark/Algorithm	RMSE	MAE	R^2	MAPE
BM Prog. duration	1.139	0.666	− 0.003	0.158
BM Prog. median TTD	1.169	0.681	0.017	0.168
DT	0.924 (0.824, 1.024)	0.667 (0.600, 0.726)	0.617 (0.558, 0.687)	0.155 (0.141, 0.172)
RF	0.863 (0.765, 0.963)	0.609 (0.542, 0.676)	0.666 (0.610, 0.727)	0.141 (0.125, 0.158)
SVM	0.898 (0.799, 0.999)	0.619 (0.555, 0.688)	0.638 (0.565, 0.707)	0.144 (0.126, 0.163)
MLP	0.930 (0.828, 1.034)	0.663 (0.605, 0.721)	0.612 (0.538, 0.682)	0.156 (0.140, 0.175)

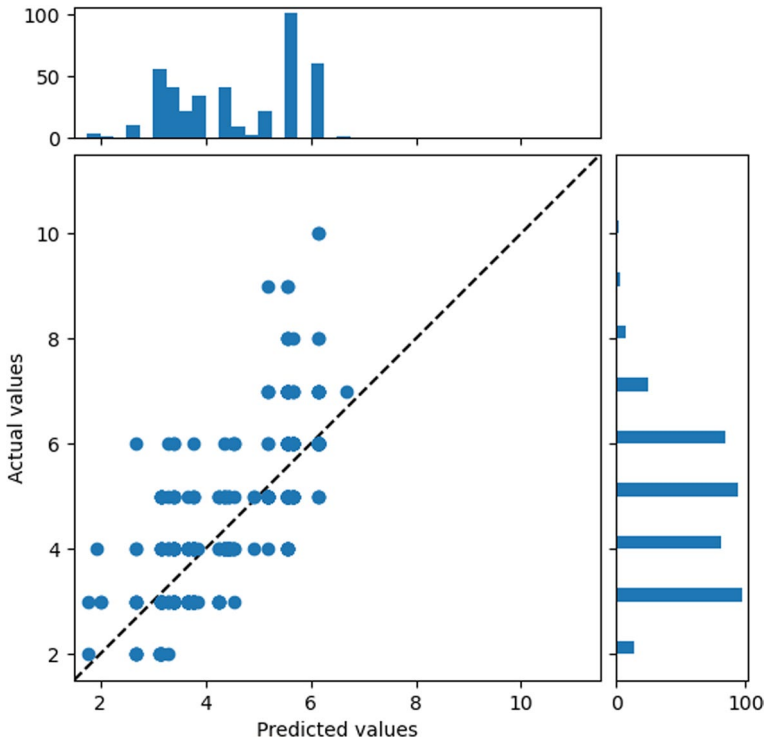


Fig. 8 Decision tree regression predictions

show that the coefficient of determination is close to zero, demonstrating a non-existent correlation between the predicted values and the actual ones. Nevertheless, the MAE is relatively small, with a maximum value of 0.666 years, which means that the mean deviation between the predicted values and the actual ones is less than an academic year. These findings underline the need for advanced models to predict the TTD of transfer students.

Figures 8 and 9 show the predicted values of TTD for the DT and RF algorithms considered in this study. The predictions for the other models are presented in Appendix D. The models were obtained with algorithms implementation provided by the scikit-learn library for Python. The performance metrics of the four models are presented in Table 3, which includes the mean and the 95% confidence interval. Regarding the performance metrics, the first conclusion is that all models perform

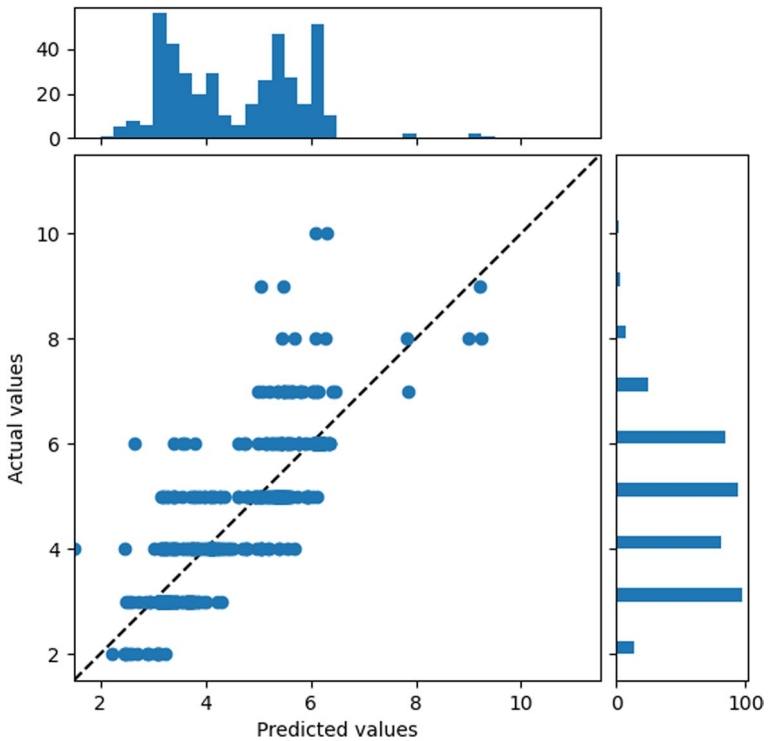


Fig. 9 Random forest regression predictions

better than the best benchmark prediction. This is evident in all metrics but is most apparent in the coefficient of determination, with results above or close to 0.6. This means that the proposed models have a high potential to support higher education decision-makers, such as programme directors. Regarding the second research question, the trained machine learning models were able to use the variables characterising the students' previous trajectories to predict the TTDs with a smaller error than the used benchmarks. This could be anticipated, as machine learning models are able to recognize complex patterns and identify interactions between features. Moreover, the use of a set of explanatory variables enables machine learning models to generalise better to unseen data.

The RF model performed the best according to all metrics. It had the lowest deviation between the predicted values and the actual ones while having the best correlation. The DT model had the worst performance metrics, even though the difference when compared to the other models was small. By analysing the predictions of each model, it is possible to identify some trends. The estimates from the DT model are translated into a cloud of points with a reduced range of values. The model could not predict TTD values higher than seven, even though they represent over three per cent of the test dataset. It also struggled to estimate values in the lower range, predicting values below the actual ones. In general, the DT model underpredicted the TTD of students with complex trajectories, since most of the points in the cloud are located above the diagonal line. The

histogram of predicted values shows two large columns representing TTDs of five and six years, indicating a better performance for integrated master's programmes.

Figure 9 shows the TTD values predicted by the RF model against the actual values, which are translated into a set of points that are close to the diagonal line. The predicted values cover a more extensive range, though the model did not catch the TTD of ten years. In the lower range, the cloud of points in Fig. 9 is also in line with the expected values. The points are scattered around the diagonal line in a smaller range, with a balanced number of points above and below. The histogram of predicted values shows three peaks, corresponding approximately to the values of three, five, and six years. These coincide with the three highest peaks of actual values.

The scatter plots of the SVM and the MLP are very similar. The scattering of points around the diagonal lines is wider in both these models when compared against the RF model. They share the same difficulties as the other models in trying to predict the TTD for large values. In terms of the histogram of the predicted values, they show a more continuous distribution of values. Although the most relevant peaks are still present, there is a quasi-continuous distribution of values in the histogram. The SVM model tends to underpredict the TTDs, especially for the lower values, while the MLP model tends to overpredict the TTDs.

Discussion

In line with several studies on education data mining that focus on RF models (Martins et al., 2019), the results obtained in this study highlight the potential of this machine learning technique. However, although RF models tend to outperform other machine learning algorithms, their results are difficult to interpret due to the stochastic nature of the decision path. The RF model combines the predictions of randomly generated tree predictors and uses the ensemble's variability to produce a more robust prediction (Breiman, 2001). Since RF models combine multiple DTs, there is no single decision path, making the model less interpretable. On the other hand, DTs, as the one shown in Fig. 11, are explainable models where the decision processes are easy to track (Quinlan, 1993). In the very first level, the median TTD of the programme is used for splitting the observations, revealing that this is the most promising feature to discriminate TTDs. This may mean that students tend to follow a similar pattern to other transfer students,

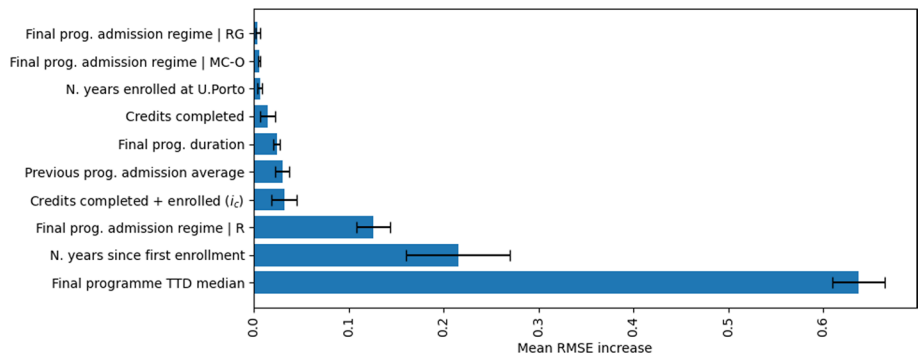


Fig. 10 Feature importance using permutation on the full model

although this past trend was not enough to estimate TTDs. Indeed, students usually transferred from the same programmes. In the second level, the decision was based on the final programme admission regime and the number of years since the first enrolment, revealing that the way the students were admitted and the time taken to transfer between programmes (perhaps due to potential credit transfers) impacted their TTDs. The further down we go in the branches, the more features are used for the decision process.

Figure 10 shows the increase in the RMSE for the top ten features, i.e. the importance of the variables, computed for the RF model. In line with the results provided by the DT, the graph shows that the median TTD for the final programme is the feature that affects the RMSE the most. This result helps explain why the benchmarks used provided reasonable predictions even though they did not take into account any other information about the student. As mentioned before, this may occur because most students complete programmes following the pattern of their colleagues from other years. This is clearly shown in the distribution of TTDs in Figs. 5 and 6, where the peak durations for bachelor's (three years) and integrated master's (five or six years) degrees are visible. The second most relevant feature refers to the period in years from the first enrolment in the university to the year of the programme change. This variable can be used by programme directors to identify students who will have more difficulties in graduating within the expected time frame. The final programme admission regime, i.e. re-enrolment (R), is also relevant in the model. This may show that students who re-enrol after a break in their studies may be following a different academic path than those who transfer to another programme without a break in their studies. The variable representing the number of credits completed before the programme change together with the credits enrolled at the beginning of the year of change also plays an important role in the model. This may be connected with the possibility of getting credits from the ECTS credits already completed. The admission average of the previous programme closes the top five most influential features. This feature belongs to the class of academic data and is often referred to in the literature as a strong predictor of academic success (Miguéis et al., 2018).

This model inspection technique shows that from the original 30 variables, a few dominate the model's performance, answering the last research question proposed. Programme directors may also look at the patterns of previous transfer students regarding subject choices and the sequence of these choices in order to identify opportunities to reduce TTDs or to guide new transfer students.

Overall, higher education institutions can benefit from accurately estimating TTDs, particularly those of students who have changed programmes. Identifying students at risk of taking longer to complete their degrees enables the promotion of early intervention and support measures to help them stay on track. This may include support services such as tutoring, mentoring, and career counseling to help students achieve their academic goals more effectively (Brock, 2010). In this way, the estimation of TTDs can enable resources to be allocated more efficiently by better understanding when students are likely to need additional academic support. In addition, TTD predictions can enable institutions to offer students alternative graduation trajectories that may be more

appropriate for each student, thus encouraging shorter academic trajectories (Sidebotham et al., 2015).

Study limitations

The quality of the present research was limited by several factors, namely the quantity and quality of the available data. The number of transfer students was not large and their characterisation was rather limited, with missing values for some variables. Another limitation of the present study relates to the granularity of the data and the moment in which the data was collected. The available data refers to the beginning of the academic years, although data collected at the beginning of each semester would have been more appropriate. For example, the number of credits enrolled and approved differs for each semester of academic study and it may have been beneficial to consider this difference in the model.

Moreover, for each academic year, it was only possible to obtain the number of credits enrolled, the cumulative number of credits approved in the programme, and the students' GPA for a given programme. The number of credits completed in a given year was not directly available. This number had to be computed based on the difference in the cumulative credits that students had in two consecutive years. However, when a student changes programme, it is impossible to compute the number of credits completed in the year before the change. Thus, regarding the credits completed, we assumed the estimate of the total number of credits completed before Y_r plus the number of credits enrolled in Y_r. In addition, a student's GPA is only known at the beginning of an academic year. Since most students change programme after the first year, the GPA of the previous programme at the moment of transfer is not known in most cases. For this reason, this variable, which the literature highlights as a significant predictor of student success (Berzenski, 2019; Iatrellis et al., 2020), was not included in the proposed models.

Most machine learning algorithms cannot handle missing values. For this reason, only 2047 of the original 2743 samples of students with complex trajectories were used to train the models. Not only does this result in a smaller dataset, potentially reducing the performance of the models, but it also reduces the diversity captured by the variables considered in the models. For example, in the time horizon considered in this study, there were transfer students who were not Portuguese. However, by not considering students with missing values, these students were excluded from the analysis and, consequently, the variable capturing the nationality was also excluded, due to the lack of different values. This can affect the quality of the models and prevent them from giving good results for situations not covered in the dataset used in the study. It should be noted that we chose not to impute missing values because of the possibility of introducing bias. In addition, imputing missing data using the wrong approach may lead to incorrect model predictions.

Some algorithms can handle nominal variables directly, as is the case of DTs. However, the scikit-learn (Buitinck et al., 2013; Pedregosa et al., 2011) algorithmic implementation adopted in this study only accepts numerical features. For this reason, all nominal variables had to be one-hot encoded. Some variables had a large cardinality, such as the

programme name (53 different values). This means that the transformed dataset used to train the models was sparse. Not only does this increase the computational time of the training phase, but it may also reduce the performance of the models.

In addition, some variables usually explored in studies related to dropping out and student academic performance were either missing or unknown in this particular case study. This was the case of parents' education and place of residence, which, according to Aparicio-Chueca et al. (2019) and other researchers, play an essential role in predictive models.

Finally, the present study was limited in geographical scope, as it only considered student transfers within the University of Porto. Although the use of unseen data and the cross-validation approach used to test the performance of the models guaranteed that the models were able to generalise in the present context, we cannot assume that the models have transferability. Indeed, in order to assess the ability of the models to generalise to other contexts, we would need to validate them on data from other contexts.

Conclusions and future work

In this study, the TTD of higher education students with complex trajectories, i.e. including a programme transfer, was characterised and predicted using machine learning models. For this purpose, we used a dataset composed of students from the University of Porto whose first enrolment occurred between 2005 and 2015 and who were tracked until 2020. The dataset included 2047 students with complex trajectories, i.e. those who changed programme during the analysis period, for which it was possible to characterise the TTD.

Our analysis demonstrated that the TTD of students with complex trajectories is statistically different from those without a complex trajectory. While students with complex trajectories graduate faster at the bachelor's degree level, the opposite is true at the integrated master's degree level. This reinforces the need to provide decision-makers, namely programme directors, with a tool that allows them to anticipate how long a student who has just enrolled in a programme will take to complete it after a transfer.

Four machine learning algorithms were used to predict the TTD of students with complex trajectories. The results revealed that predictive modelling is effective in the academic domain, particularly in predicting TTD in complex scenarios, and that decision-makers can use such models to plan institutional actions and optimise their limited resource allocation. By accurately predicting when students are likely to complete their degree programmes, institutions can take proactive measures to enhance students' academic experience and improve overall educational outcomes. Once students at risk of taking longer to complete their studies have been identified, advisers can use the model's insights to provide personalised advice and support. This could include creating tailored academic plans, suggesting appropriate courses, or referring students to support services. Institutions can also provide additional tutoring and mentoring for students who are likely to take longer to complete their studies.

The RF model had the best performance out of the four models, while the DT model performed the worst. Yet, all four models performed better regarding the goodness-of-fit metrics than the two benchmark models. While the RF model showed a better prediction capacity than the other models, similar to that of neural networks, it is an opaque or black box model. These models are difficult to understand because the predictions are based on a decision process that is not understandable by humans. Thus, we conducted a feature importance determination using permutation on the model to help identify the variables that affect the model's performance the most. We concluded that the most relevant factors to predict the TTD of students with complex trajectories were the median TTD of the final programme, the number of years since the first enrolment, and the admission regime of the previous programme.

Although the DT model performed the worst, it is an explainable model where it is possible to interpret the decision process. The DT model showed in the first branches the same variables highlighted in the variables' importance analysis: the median TTD, the final programme admission regime, and the number of years since the first enrolment. Thus, it is possible to state that these are the most relevant factors for predicting TTDs.

The choice between model transparency and predictive accuracy is a critical consideration when developing machine learning models such as RFs and DTs. These two aspects often represent a trade-off, and the decision can have a significant impact on a model's acceptance and usefulness in real-world scenarios. If a model's predictions are to be used in contexts where explicability is crucial (e.g. healthcare, finance, or law), an interpretable model may be preferable, even if it sacrifices some prediction accuracy. Conversely, if prediction accuracy is paramount, models such as RFs may be the better choice. The decision should be based on the specific requirements and constraints of the scenario in which the model will be used.

From an educational point of view, we believe that the proposed model is relevant, as it helps to predict the TTD of transfer students and can open possibilities such as more timely interventions from decision-makers that may lead to TTD reductions and improvements in the quality of students' academic experiences.

Regarding future work, the models could be further improved by enriching the dataset with more data. In particular, some observations were not included in the dataset due to missing values in some features. This primarily affected variables related to the period preceding the programme change, like the admission average or the application preference for the previous programme. This exclusion of observations affected other features, like nationality, which initially included several countries and, in the final dataset, resulted in a cohort of students of Portuguese nationality. This may limit the potential of the predictive models proposed.

We also believe that it would be relevant for future works to develop a qualitative study, targeting transfer students, to assess what motivates the delays in their trajectory after changing programme. Identifying these factors may lead to the development of richer models that accommodate other relevant predictive variables, in addition to those covered in this study.

Appendix A

Decision tree plot (Fig. 11).

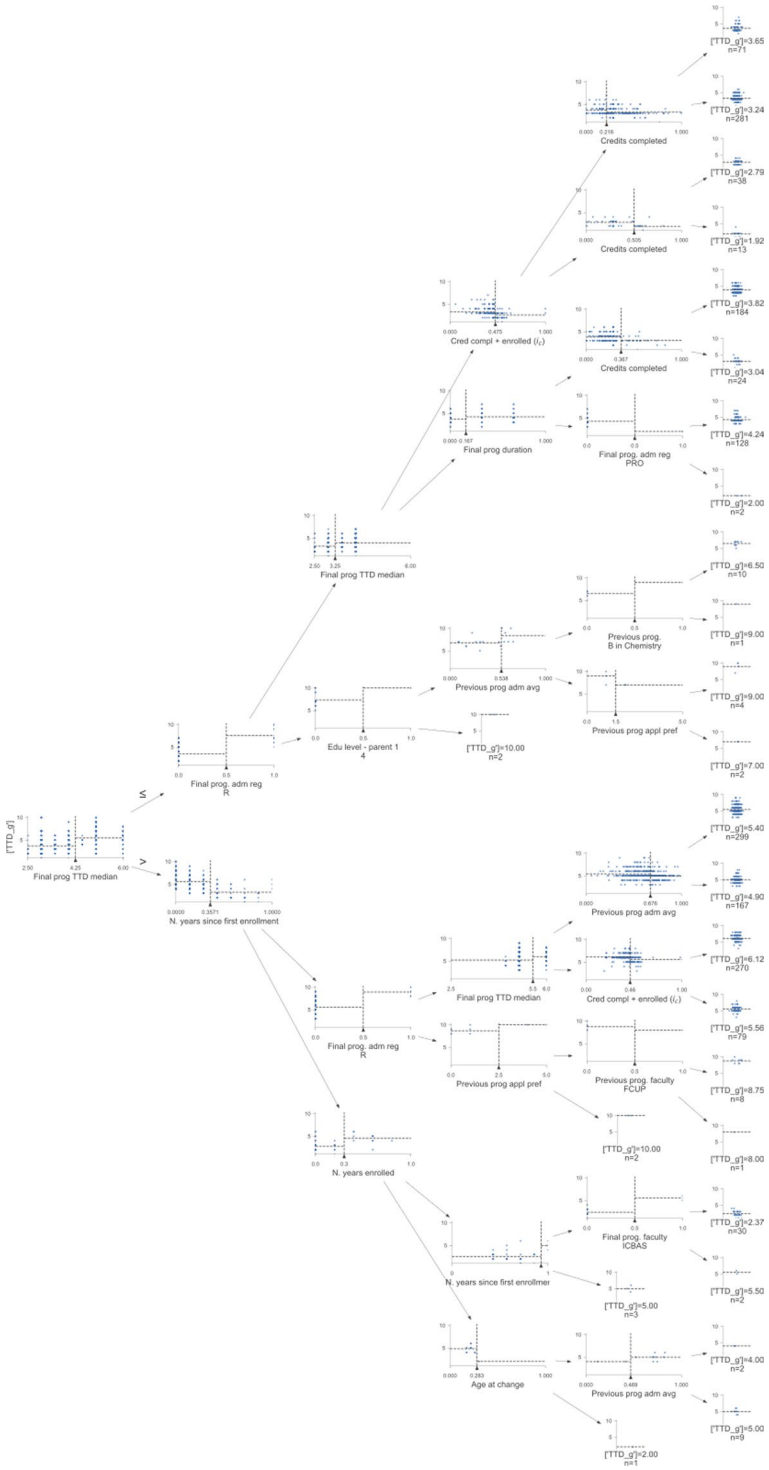


Fig. 11 Decision tree model

Appendix B

See Table 4

Table 4 Features used in the study

#	Attribute	Period	Type	Values (frequency)/Mean (std. deviation)
1	Gender	h_s	D	Female: 53.38% Male: 46.62%
2	High school type	h_s	PAD	Public: 66.8% Private: 27.67% Public and private: 5.04% Unknown: 0.49%
3	Education level of parent 1	Y_p	D	Unknown: 55.93% Higher education—Bachelor's degree: 12.29% Secondary education—Year 12 or equivalent: 10.04% Third stage of basic education—Year 9: 6.12% First stage of basic education—Year 4: 3.77% (...)
4	Education level of parent 2	Y_p	D	Unknown: 55.34% Higher education—Bachelor's degree: 15.77% Secondary education—Year 12 or equivalent: 8.52% Third stage of basic education—Year 9: 5.93% Master's degree: 3.28% (...)
5	% of time working	Y_p	D	0.01 (0.09)
6	% of time displaced	Y_p	D	0.27 (0.44)
7	% of time with scholarship	Y_p	D	0.09 (0.28)
8	Credits completed	Y_p	AD	59.6 (30.78)
9	N. enrolments	Y_p	AD	2.43 (0.8)
10	N. years enrolled at U.Porto	Y_p	AD	2.34 (0.74)
11	N. prog.	Y_p	AD	2.05 (0.23)
12	N. prog. with credits enrolled	Y_p	AD	1.98 (0.25)
13	Credits completed + enrolled (i_c)	Y_p	AD	57.81 (10.45)
14	N. years since first enrolment	Y_p	AD	2.56 (1.1)
15	Previous prog. admission regime	Y_r	PAD	General contingent: 51.57% General regime: 47.06% Special contingents: Madeira: 0.59% Special contingents: Azores: 0.44% Special contingents: Portuguese and family emigrants: 0.15% (...)
16	Previous prog. application preference	Y_r	PAD	1: 36.88% 2: 26.1% 3: 14.5% 4: 9.35% 5: 7.0% (...)
17	Previous prog. admission average	Y_r	PAD	16.22 (1.71)
18	Previous prog.	Y_r	I	IM Elect and Comp Eng: 7.35% IM Dental Medicine: 6.27% IM in Mech Eng: 5.68% IM in Pharm Sciences: 5.53% IM in Civil Engineering: 5.0% (...)
19	Previous prog. degree	Y_r	I	IM: 57.79% B: 42.21%
20	Previous prog. faculty	Y_r	I	FEUP: 33.59% FCUP: 19.29% FLUP: 11.31% ICBAS: 6.42% FMDUP: 6.27% (...)
21	Working	i_c	D	no: 98.04% yes: 1.96%
22	Displaced	i_c	D	no: 72.48% yes: 27.52%
23	Scholarship	i_c	D	no: 89.86% yes: 10.14%
24	Age at change	i_c	D	19.47 (1.51)
25	Final prog. admission regime	i_c	PAD	General regime: 26.15% General contingent: 25.37% Other course change: 12.39% Course Change 1st Year 1st Semester: 11.26% Change of course: 6.56% (...)
26	Final prog. faculty	i_c	I	FEUP: 24.93% FLUP: 17.29% ICBAS: 14.05% FCUP: 12.44% FEP: 7.35% (...)
27	Final prog.	i_c	I	IM in Medicine: 17.14% IM in Eng & Ind Mngt: 5.0% IM in Mech Engineering: 4.75% B in Comm Sci: J, PR, M: 4.36% IM in Informatics and Computing Engineering: 4.11% (...)
28	Final prog. degree	i_c	I	IM: 53.72% B: 46.28%
29	Final prog. duration	i_c	I	4.33 (1.17)
30	Final programme TTD median	i_c	I	4.48 (1.13)

D—Demographic; PAD—Prior academic data; I—Institutional; AD—Academic data Sghir et al. (2022)

Appendix C

See Table 5

Table 5 Number of transfers and top-3 programmes of destination

Previous Prog.	#	Final Prog.#1	Perc.	Final Prog.#2	Perc.	Final Prog.#3	Perc.
IM Elect & Comp Eng	157	IM Inf & Comp Eng	30.6%	IM Mech Eng	14.0%	B Business Adm	7.0%
IM Dental Med	144	IM Med	95.1%	B Economics	1.4%	B Comm Design	0.7%
IM Mech Eng	121	IM Eng & Ind Mgmt	43.0%	B Economics	6.6%	B Business Adm	5.8%
IM Pharm Sci	120	IM Medicine	42.5%	IM Dental Medicine	15.0%	B Nutrition Sci	6.7%
IM Vet Med	107	IM Medicine	73.8%	IM Dental Med	7.5%	IM Pharm Sci	3.7%
IM Civil Eng	107	IM Mech Eng	35.5%	IM Elect & Comp Eng	13.1%	B Economics	9.3%
IM Chemical Eng	91	IM Pharm Sci	18.7%	IM BioEng	17.6%	IM Eng & Ind Mgmt	9.9%
B Biology	89	B Nutrition Sci	22.5%	IM Vet Medicine	11.2%	IM Pharm Sci	11.2%
IM Inf & Comp Eng	72	IM Elect & Comp Eng	11.1%	IM Eng & Ind Mgmt	11.1%	B Economics	9.7%
B Lang, Lit & Cult	70	B Applied Lang	32.9%	B Lang & Int Rel	27.1%	B Comm Sci: J, PR, M	11.4%
B Biochemistry	66	IM Pharm Sci	22.7%	B Nutrition Sci	15.2%	IM Medicine	12.1%
B Law	65	B Comm Sci: J, PR, M	23.1%	B Lang, Lit & Cult	16.9%	B History	10.8%
IM BioEng	64	IM Medicine	53.1%	IM Mech Eng	7.8%	IM Eng & Ind Mgmt	6.2%
IM Physical Eng	62	B Physics	27.4%	B Chemistry	11.3%	IM Elect & Comp Eng	8.1%
IM Env Eng	56	IM Chemical Eng	26.8%	B Business Adm	7.1%	IM Eng & Ind Mgmt	7.1%
IM Net & Inf Sys Eng	52	B Comp Sci	30.8%	IM Inf & Comp Eng	15.4%	B Eng Sci	9.6%
B Fine Arts	51	B Comm Design	60.8%	IM Architecture	13.7%	B Philosophy	3.9%
B Economics	48	B Business Adm	31.2%	B Law	12.5%	B Comm Sci: J, PR, M	10.4%
B Physics	46	IM Physical Eng	19.6%	B Comp Sci	10.9%	IM Elect & Comp Eng	8.7%
IM Medicine	43	IM Medicine	32.6%	IM Eng & Ind Mgmt	11.6%	IM CiviB Eng	7.0%
B Nutrition Sci	43	IM Medicine	48.8%	IM Pharm Sci	23.3%	IM Dental Medicine	9.3%
B Chemistry	41	B Biology	26.8%	B Biochemistry	17.1%	B Env Sci & Tech	14.6%
Previous Prog.	**#**	**Final Prog.#1**	**Perc.**	**Final Prog.#2**	**Perc.**	**Final Prog.#3**	**Perc.**
B Env Sci & Tech	37	B Biology	43.2%	IM Env Eng	16.2%	B Biochemistry	8.1%
IM Architecture	35	B Fine Arts	17.1%	B Comm Design	17.1%	IM Civil Eng	14.3%
B Aquatic Sci	32	IM Vet Medicine	40.6%	B Biology	15.6%	IM Medicine	6.2%
B Mathematics	30	B Comp Sci	13.3%	B Eng Sci	10.0%	B Biology	10.0%
B Criminology	30	B Law	43.3%	B Comm Sci: J, PR, M	13.3%	IM Psychology	10.0%
B Eng Sci	29	IM Elect & Comp Eng	20.7%	B Biology	20.7%	B Mining & Geo-Env Eng	10.3%
IM Eng & Ind Mgmt	28	IM Medicine	21.4%	IM Mech Eng	17.9%	B Economics	14.3%
B Astronomy	27	B Eng Sci	22.2%	B Biology	18.5%	B Geology	11.1%
B Edu Sci	27	IM Psychology	88.9%	B Sociology	3.7%	B Sports Sci	3.7%
IM Psychology	26	B Nutrition Sci	15.4%	B Comm Sci: J, PR, M	15.4%	B of Arts in Archaeology	7.7%
B of Arts in Info Sci	25	B Comm Sci: J, PR, M	52.0%	B Lang, Lit & Cult	12.0%	B Edu. Sci	8.0%
IM Metall & Mat Eng	23	IM Mech Eng	43.5%	IM Elect & Comp Eng	17.4%	IM CiviB Eng	8.7%
B Comp Sci	22	IM Net. & Inf Sys Eng	36.4%	B Env. Sci & Tech.	13.6%	B Sports Sci	9.1%
B History	22	B Comm Sci: J, PR, M	27.3%	B Law	9.1%	B Geography	9.1%
B Comm Design	21	IM Architecture	33.3%	B Fine Arts	28.6%	IM Psychology	9.5%
B Philosophy	21	B Sociology	23.8%	B of Arts in Info. Sci	14.3%	B History	14.3%
B Comm Sci: J, PR, M	21	B Lang, Lit & Cult	14.3%	B Lang & Int Rel.	14.3%	B Applied Lang	9.5%

Table 5 (continued)

Previous Prog.	#	Final Prog.#1	Perc.	Final Prog.#2	Perc.	Final Prog.#3	Perc.
B Mining & Geo-Env. Eng	19	IM CiviB Eng	26.3%	IM Elect & Comp Eng	15.8%	IM Env. Eng	10.5%
B Lang & Int Rel.	18	B Lang, Lit & Cult	27.8%	B Applied Lang	11.1%	B Law	11.1%
B Lang Sci	18	B Comm Sci: J, PR, M	33.3%	B Applied Lang	16.7%	B Lang, Lit & Cult	16.7%
B Geography	18	B Sociology	16.7%	B of Arts in Archaeology	16.7%	B History	16.7%
B Sociology	15	IM Psychology	46.7%	B Lang & Int Rel.	13.3%	B Comm Sci: J, PR, M	13.3%
B PT Studies	15	B Lang, Lit & Cult	26.7%	B Philosophy	13.3%	IM Psychology	13.3%
B Geology	15	B Biology	40.0%	B Land. Arch.	20.0%	B Env. Sci & Tech.	13.3%
B Business Adm	13	B Economics	46.2%	B PT Studies	7.7%	IM Psychology	7.7%
B History of Art	12	B Fine Arts	16.7%	B Applied Lang	16.7%	B of Arts in Info. Sci	16.7%
B of Arts in Archaeology	11	B History	27.3%	B Lang, Lit & Cult	18.2%	B Comm Sci: J, PR, M	18.2%
B Land. Arch.	10	B Eng Sci	30.0%	B Comm Design	10.0%	B of Arts in Info. Sci	10.0%
B Applied Lang	8	B Lang & Int Rel.	75.0%	B Lang, Lit & Cult	25.0%	B Agricultural Eng	0.0%
B Sports Sci	4	B Biology	25.0%	IM Medicine	25.0%	IM Inf & Comp Eng	25.0%

Appendix D

SVM and MLP regression plots (Figs. 12, 13).

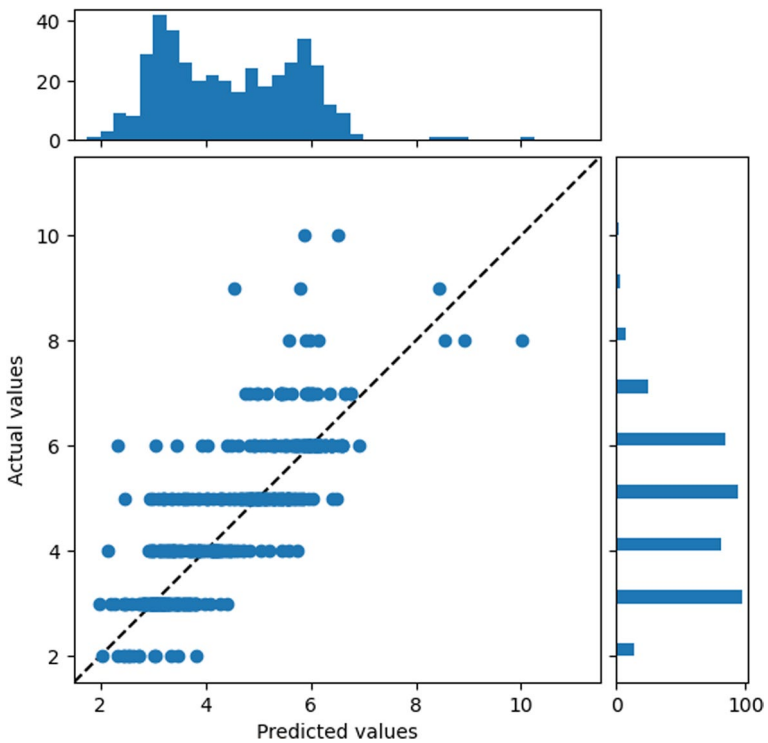
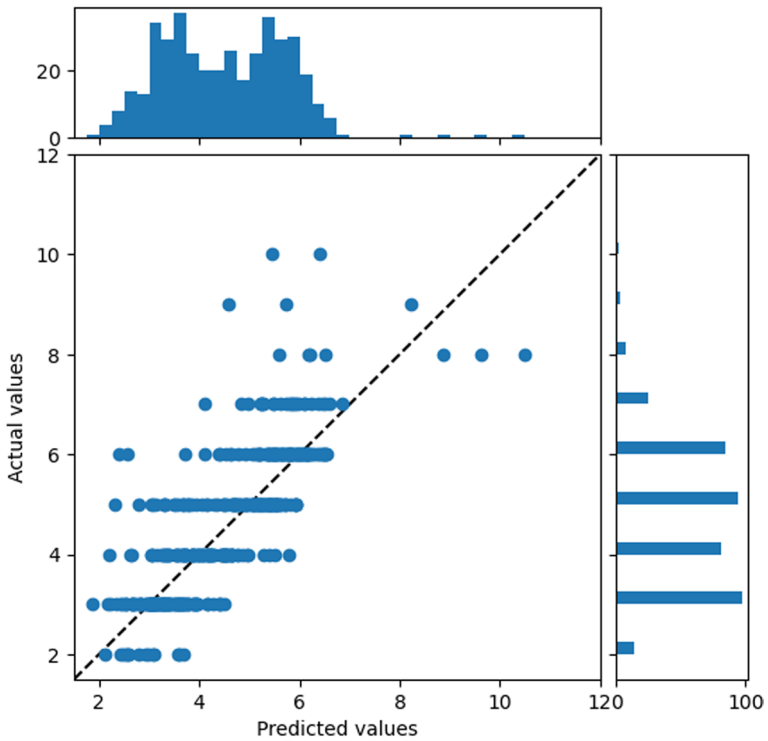


Fig. 12 Support vector machine regression predictions



ECTS - European Credit Transfer and Accumulation System
DT - Decision tree
RF - Random forest
MAE - Mean absolute error
MAPE - Mean absolute percentage error
MLP - Multilayer perceptron
RMSE - Root mean square error
SVM - Support vector machine
TTD - Time to degree

Fig. 13 Multilayer preceptron regression predictions

Abbreviations

ECTS	European Credit Transfer and Accumulation System
DT	Decision tree
RF	Random forest
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MLP	Multilayer perceptron
RMSE	Root mean square error
SVM	Support vector machine
TTD	Time to degree

Acknowledgements
Not applicable.

Author contributions
JPP contributed to the pre-processing of the data, the training and optimization of the machine learning models, produced the graphics and analysed the results of the model. VM contributed to the state-of-the-art, introduction, and methodology and analyzed the students' complex trajetories. AS contributed to the conceptualization of the study, state-of-the-art, introduction, methodology and final review of the article.

Funding
This work was funded by the European Union through the ERASMUS+ project with reference 2020-1-ES01-KA203-082842 and co-supported through strategic funding from FCT UIDB044232020 and UIDP044232020.

Availability of data and materials
The data that supports the findings of this study is available from the University of Porto's Rectory but restrictions apply to the availability of this data, which was used under license for the current study, and so is not publicly available. The data can however be made available by the authors upon reasonable request and with the permission of the University of Porto's Rectory.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 23 June 2023 Accepted: 8 January 2024
Published online: 29 January 2024

References

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics, 37*, 13–49. https://doi.org/10.1016/j.tele.2019.01.007

Almeida, L.S., Castro, R.V.d., Seminário. (2016). "Ser Estudante no Ensino Superior: O caso dos estudantes do 1° ano", Ser estudante no ensino superior: o caso dos estudantes do 1° ano. Universidade do Minho. Centro de Investigação em Educação (CIEd), Portugal. Accepted: 2016-06-21T15:41:53Z. http://repositorium.sdum.uminho.pt/ Accessed 2023-04-24

Aparicio-Chueca, P., Domínguez-Amorós, M., & Maestro-Yarza, I. (2019). Beyond university dropout. An approach to university transfer. *Studies in Higher Education, 46*(3), 473–484. https://doi.org/10.1080/03075079.2019.1640671

Barakat, B., & Shields, R. (2019). Just another level? Comparing quantitative patterns of global expansion of school and higher education attainment. *Demography, 56*(3), 917–934. https://doi.org/10.1007/s13524-019-00775-5

Basavaraj, P., Garibay, I. (2019). Dropout vs. Time to degree. In: Proceedings of the 20th Annual SIG Conference on Information Technology Education, p. 154. ACM, Tacoma, WA, USA. https://doi.org/10.1145/3349266.3351374

Berger, J. B., & Braxton, J. M. (1998). Revising Tinto's interactionalist theory of student departure through theory elaboration: Examining the Role of Organizational Attributes in the Persistence Process. *Research in Higher Education, 39*(2), 103–119. https://doi.org/10.1023/A:1018760513769

Berzenski, S. R. (2019). The when and who of graduation and dropout predictors: A moderated hazard analysis. *Journal of College Student Retention: Research, Theory & Practice, 23*(3), 768–792. https://doi.org/10.1177/1521025119875104

Berzenski, S. R. (2021). The when and who of graduation and dropout predictors: A moderated hazard analysis. *Journal of College Student Retention: Research, Theory & Practice, 23*(3), 768–792. https://doi.org/10.1177/1521025119875104

Bhaskaran, S. S., Lu, K., & Aali, M. A. (2017). Student performance and time-to-degree analysis by the study of course-taking patterns using J48 decision tree algorithm. *International Journal of Modelling in Operations Management, 6*(3), 194–213. https://doi.org/10.1504/IJMOM.2017.084814

Biffl, G., & Isaac, J. (2002). Should higher education students pay tuition fees? *European Journal of Education, 37*(4), 433–455.

Boegeholz, R., Guerra, J., & Scheihing, E. (2022). Exploring risk of delay in academic trajectories in two undergraduate programs. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 17*(3), 290–300. https://doi.org/10.1109/rita.2022.3191298

Boehmke, B., & Greenwell, B. (2019). Random forests. In: Hands-on Machine Learning with R, pp. 203–219. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9780367816377-11

Bowman, N. A., & Holmes, J. M. (2018). Getting off to a good start? First-year undergraduate research experiences and student outcomes. *Higher Education, 76*(1), 17–33.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Brezavšcek, A., Bach, M. P., & Baggia, A. (2017). Markov Analysis of Students' Performance and Academic Progress in Higher Education. *Organizacija, 50*(2), 83–95. https://doi.org/10.1515/orga-2017-0006

Brock, T. (2010). Young adults and higher education: Barriers and breakthroughs to success. *The Future of Children, 20*(1), 109–132. https://doi.org/10.1353/foc.0.0040

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122

Carreira, P., & Lopes, A. S. (2019). Drivers of academic pathways in higher education: Traditional vs. non-traditional students. *Studies in Higher Education, 46*(7), 1340–1355. https://doi.org/10.1080/03075079.2019.1675621

Casanova, J.R., & Almeida, L.S. (2016). Diversidade de públicos no Ensino Superior: Antecipando riscos na qualidade da adaptação e do sucesso académico em estudantes do 1.° ano. Accepted: 2021-04-14T10:34:43Z Publisher: Edições ISPGaya. Accessed 2023-04-24

Casuat, C.D., & Festijo, E.D. (2019). Predicting Students' Employability using Machine Learning Approach. In: 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), pp. 1–5. https://doi.org/10.1109/ICETAS48360.2019.9117338

Chin-Newman, C. S., & Shaw, S. T. (2013). The anxiety of change: How new transfer students overcome challenges. *Journal of College Admission, 221*, 14–21.

Cortez, P., & Silva, A.M.G. (2008). Using data mining to predict secondary school student performance. EUROSIS-ETI.
Czajkowski, M., & Kretowski, M. (2016). The role of decision tree representation in regression problems—an evolutionary perspective. *Applied Soft Computing, 48*, 458–475. https://doi.org/10.1016/j.asoc.2016.07.007
Dias, D. (2015). Has massification of higher education led to more equity? Clues to a reflection on Portuguese education arena. *International Journal of Inclusive Education, 19*(2), 103–120. https://doi.org/10.1080/13603116.2013.788221
Ferrão, M.E., & Almeida, L.S. (2019). Student's access and performance in the Portuguese Higher Education: Issues of gender, age, socio-cultural background, expectations, and program choice. Avaliação: Revista da Avaliação da Educação Superior (Campinas) 24, 434–450. https://doi.org/10.1590/S1414-40772019000200006. Publisher: Publicação da Rede de Avaliação Institucional da Educação Superior (RAIES), da Universidade Estadual de Campinas (UNICAMP) e da Universidade de Sorocaba (UNISO). Accessed 2023-04-28
Ferrão, M.E., Almeida, L.S. (2018). Multilevel modeling of persistence in higher education. Ensaio: Avaliação e Políticas Públicas em Educação 26(100), 664–683. Accessed 2023-04-24
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1–39. https://doi.org/10.1214/aos/1013203451
Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367–378. https://doi.org/10.1016/s0167-9473(01)00065-2
Giani, M. S. (2015). The postsecondary resource trinity model: Exploring the interaction between socioeconomic, academic, and institutional resources. *Research in Higher Education, 56*(2), 105–126. https://doi.org/10.1007/s11162-014-9357-4
Goldrick-Rab, S. (2006). Following their every move: An investigation of social-class differences in college pathways. *Sociology of Education, 79*(1), 61–79.
Goma, H. (2023). The Need To Investigate Complex Trajectories—ETHE Journal Blog. http://etheblog.com/2023/03/18/the-need-to-investigate-complex-trajectori es/ Accessed 2023-04-21
Haas, C., & Hadjar, A. (2020). Students' trajectories through higher education: A review of quantitative research. *Higher Education, 79*(6), 1099–1118. https://doi.org/10.1007/s10734-019-00458-5
Hadjar, A., & Becker, R. (2009). Expected and Unexpected Consequences of the Educational Expansion in Europe and the US Theoretical Approaches and Empirical Findings in Comparative perspective. Haupt Verlag, Switzerland. https://orbilu.uni.lu/handle/10993/1899 Accessed 2023-04-28
Hailikari, T., Sund, R., Haarala-Muhonen, A., & Lindblom-Ylänne, S. (2019). Using individual study profiles of first-year students in two different disciplines to predict graduation time. *Studies in Higher Education, 45*(12), 2604–2618. https://doi.org/10.1080/03075079.2019.1623771
Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020). Student performance prediction model based on supervised machine learning algorithms. *IOP Conference Series: Materials Science and Engineering, 928*(3), 032019. https://doi.org/10.1088/1757-899x/928/3/032019
Hovdhaugen, E. (2009). Transfer and dropout: Different forms of student departure in Norway. *Studies in Higher Education, 34*(1), 1–17. https://doi.org/10.1080/03075070802457009
Iatrellis, O., Savvas, I., Fitsilis, P., & Gerogiannis, V. C. (2020). A two-phase machine learning approach for predicting student outcomes. *Education and Information Technologies, 26*(1), 69–88. https://doi.org/10.1007/s10639-020-10260-x
Iatrellis, O., Savvas, I. K., Kameas, A., & Fitsilis, P. (2020). Integrated learning pathways in higher education: A framework enhanced with machine learning and semantics. *Education and Information Technologies, 25*(4), 3109–3129. https://doi.org/10.1007/s10639-020-10105-7
Karalar, H., Kapucu, C., & Gürüler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of Educational Technology in Higher Education, 18*(1), 63. https://doi.org/10.1186/s41239-021-00300-y
MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1: Statistics, pp. 281–297. University of California Press.
Martins, M.P.G., Migueis, V.L., & Fonseca, D.S.B. (2018). Educational data mining: A literature review, 2018: 1–6. https://doi.org/10.23919/CISTI.2018.8399281
Martins, M. P. G., Miguéis, V. L., Fonseca, D. S. B., & Alves, A. (2019). A data mining approach for predicting academic success—A case study. *Advances in Intelligent Systems and Computing, 918*, 45–56. https://doi.org/10.1007/978-3-030-11890-7
Martins, M. P. G., Miguéis, V. L., Fonseca, D. S. B., & Alves, A. (2019). a data mining approach for predicting academic success—A case study. In Á. Rocha, C. Ferrás, & M. Paredes (Eds.), *Information technology and systems. Advances in intelligent systems and computing* (pp. 45–56). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-11890-7_5
Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Boosting algorithms as gradient descent. In: Solla, S., Leen, T., Müller, K. (eds.) Advances in Neural Information Processing Systems, vol. 12, pp. 1–29. MIT Press. https://proceedings.neurips.cc/paper/1999/file/96a93ba89a5b5c6c226e49b88973f46e-Paper.pdf
Miguéis, V. L., Freitas, A., Garcia, P. J. V., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36–51. https://doi.org/10.1016/j.dss.2018.09.001
Monaghan, D. B. (2019). College-going trajectories across early adulthood: An inquiry using sequence analysis. *The Journal of Higher Education, 91*(3), 402–432. https://doi.org/10.1080/00221546.2019.1647584
Moreira, J.M., Carvalho, A.C.P.L.F., & Horváth, T. (2018). A general introduction to data analytics. Wiley. https://doi.org/10.1002/9781119296294
Okun, M. A., Goegan, B., & Mitric, N. (2009). Quality of alternatives, institutional preference, and institutional commitment among first-year college students. *Educational Psychology, 29*(4), 371–383. https://doi.org/10.1080/01443410902957079

Pardo, A., Han, F., & Ellis, R. A. (2017). Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. *IEEE Transactions on Learning Technologies, 10*(1), 82–92. https://doi.org/10.1109/tlt.2016.2639508
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
Prenkaj, B., Distante, D., Faralli, S., & Velardi, P. (2021). Hidden space deep sequential risk prediction on student trajectories. *Future Generation Computer Systems, 125*, 532–543. https://doi.org/10.1016/j.future.2021.07.002
Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers Inc.
Rayner, G., & Papakonstantinou, T. (2022). The variables that predict science undergraduates' timely degree completion: A conceptual model. *Research in Science Education*. https://doi.org/10.1007/s11165-022-10064-8
Robinson, R. (2004). Pathways to completion: Patterns of progression through a university degree. *Higher Education, 47*(1), 1–20. https://doi.org/10.1023/B:HIGH.0000009803.70418.9c
Rodríguez-Gómez, D., Meneses, J., Gairín, J., Feixas, M., & Muñoz, J. L. (2016). They have gone, and now what? Understanding re-enrolment patterns in the Catalan public higher education system. *Higher Education Research & Development, 35*(4), 815–828. https://doi.org/10.1080/07294360.2015.1137886
Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery, 10*(3), 1355. https://doi.org/10.1002/widm.1355
Rosenberg, M.B., Hilton, M.L., & Dibner, K.A. (eds.) (2018). Indicators for Monitoring Undergraduate STEM Education. National Academies Press, Washington, D.C. https://doi.org/10.17226/24943. https://www.nap.edu/catalog/24943 Accessed 2023-04-21
Sánchez-Gelabert, A., Valente, R., & Duart, J. M. (2020). Profiles of online students and the impact of their university experience. *The International Review of Research in Open and Distributed Learning, 21*(3), 230–249. https://doi.org/10.19173/irrodl.v21i3.4784
Schofer, E., & Meyer, J. W. (2005). The worldwide expansion of higher education in the twentieth century. *American Sociological Review, 70*, 898–920. https://doi.org/10.1177/000312240507000602
Sghir, N., Adadi, A., & Lahmer, M. (2022). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Springer*. https://doi.org/10.1007/s10639-022-11536-0
Sidebotham, M., Fenwick, J., Carter, A., & Gamble, J. (2015). Using the Five Senses of Success framework to understand the experiences of midwifery students enroled in an undergraduate degree program. *Midwifery, 31*(1), 201–207. https://doi.org/10.1016/j.midw.2014.08.007
Sousa, L. (2021). CONJECTURES AND DEMOCRATIZATION OF HIGHER EDUCATION: THE POLYTECHNIC INSTITUTE OF VISEU CASE. EDULEARN21 Proceedings, 1911–1917. https://doi.org/10.21125/edulearn.2021.0441. Conference Name: 13th International Conference on Education and New Learning Technologies ISBN: 9788409312672 Meeting Name: 13th International Conference on Education and New Learning Technologies Place: Online Conference Publisher: IATED. Accessed 2023-04-21
Teixeira, P. N., Silva, P. L., Biscaia, R., & Sá, C. (2022). Competition and diversification in higher education: Analysing impacts on access and equity in the case of Portugal. *European Journal of Education, 57*(2), 235–254.
Terenzini, P. T., Lorang, W. G., & Pascarella, E. T. (1981). Predicting freshman persistence and voluntary dropout decisions: A replication. *Research in Higher Education, 15*(2), 109–127. https://doi.org/10.1007/BF00979592
Tieben, N. (2019). Non-completion, transfer, and dropout of traditional and non-traditional students in Germany. *Research in Higher Education, 61*(1), 117–141. https://doi.org/10.1007/s11162-019-09553-z
Tieben, N. (2020). Non-completion, transfer, and dropout of traditional and non-traditional students in Germany. *Research in Higher Education, 61*(1), 117–141. https://doi.org/10.1007/s11162-019-09553-z
Tinto, V. (1994). *Leaving college: rethinking the causes and cures of student attrition*. University of Chicago Press.
Townsend, B. K., & Wilson, K. B. (2009). The academic and social integration of persisting community college transfer students. *Journal of College Student Retention: Research, Theory & Practice, 10*(4), 405–423. https://doi.org/10.2190/CS.10.4.a
Tumen, S., Shulruf, B., & Hattie, J. (2008). Student pathways at the university: Patterns and predictors of completion. *Studies in Higher Education, 33*(3), 233–252. https://doi.org/10.1080/03075070802049145
Utter, M., & DeAngelo, L. (2015). Lateral transfer students: The role of housing in social integration and transition. *Journal of College and University Student Housing, 42*(1), 178–193.
Wang, H., Liu, H., & Zhang, X. (2016). Development trend of support vector machine and applications on the field of computer science. In: Proceedings of the 2016 International Conference on Computer Engineering, Information Science & Application Technology (ICCIA 2016). Atlantis Press. https://doi.org/10.2991/iccia-16.2016.91
Wang, Y. (2021). Who benefits more from the college expansion policy? Evidence from China. *Research in Social Stratification and Mobility, 71*, 100566. https://doi.org/10.1016/j.rssm.2020.100566
Wong, B., & Chiu, Y.-L.T. (2019). Swallow your pride and fear: The educational strategies of high-achieving non-traditional university students. *British Journal of Sociology of Education, 40*(7), 868–882. https://doi.org/10.1080/01425692.2019.1604209
Xerri, M. J., Radford, K., & Shacklock, K. (2018). Student engagement in academic activities: A social support perspective. *Higher Education, 75*(4), 589–605.
Xie, Y., Fang, M., & Shauman, K. (2015). STEM Education. *Annual Review of Sociology, 41*, 331–357. https://doi.org/10.1146/annurev-soc-071312-145659
Yi, P.-S. (2008). Institutional climate and student departure: A multinomial multilevel modeling approach. *Review of Higher Education, 31*(2), 161–183.

Publisher's Note