CESifo WORKING PAPERS

9300
2021

September 2021

Predicting Student Dropout: A Replication Study Based on Neural Networks

Jascha Buchhorn, Berthold U. Wigger

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9300*

Predicting Student Dropout: A Replication Study Based on Neural Networks

Abstract

Using neural networks, the present study replicates previous results on the prediction of student dropout obtained with decision trees and logistic regressions. For this purpose, multilayer perceptrons are trained on the same data as in the initial study. It is shown that neural networks lead to a significant improvement in the prediction of students at risk. Already after the first semester, potential dropouts can be identified with a probability of 95 percent.

Keywords: neural networks, student dropout, replication study.

Jascha Buchhorn
Karlsruhe Institute of Technology, Chair of Public Finance and Public Management
Karlsruhe / Germany
jascha.buchhorn@kit.edu

*Berthold U. Wigger**
Karlsruhe Institute of Technology, Chair of Public Finance and Public Management
Karlsruhe / Germany
berthold.wigger@kit.edu

*corresponding author

July 28, 2021

1. Introduction

Dropout is a widespread problem in the German higher education system. About 30 percent of students in Germany do not complete their studies (Heublein U. , 2014; Heublein, et al., 2017). At the same time, more knowledge-intensive production processes and demographic change are increasingly leading to a shortage of skilled workers in the German labor market and thus to an increased demand for university graduates. The shortage of skilled workers is particularly critical in the STEM subjects - these subject areas also have high dropout rates (Hetze, 2011). Dropout is therefore associated with sensitive costs not only for students and universities, but also for the development of society as a whole.

Early detection systems, as recent work has shown (Kemper, Vorhoff, & Wigger, 2020; Berens, Schneider, Gortz, Oster, & Burghoff, 2019; Ram, Wang, Currim, & Currim, 2018), can be a valuable addition to higher education institutions' efforts to detect dropout. In this regard, identifying at-risk students represents a critical first step in addressing dropout with targeted interventions such as learning assistance or mentoring programs. It enables a targeted and thus efficient use of the scarce educational and administrative resources of a university.

This is where the present paper comes in and expands on a previous approach. The starting point is the paper by Kemper et al. (2020). This paper describes the methodology and the results of a case study conducted on study progress data of the Faculty of Economics at the Karlsruhe Institute of Technology (KIT). The study shows that study success can be identified with respectable estimation accuracies of 85 percent in the first study semester and up to 95 percent in the third study semester already with relatively simple procedures and on a data basis that is not subject to data protection laws.

The paper by Kemper et al. (2020) employs decision trees and logit models for classification. The present paper aims to replicate the results of Kemper et al. (2020) using neural networks. For this purpose, multilayer perceptrons (MLP) will be trained on the same data and their prediction performance will be compared with the results of the initial study.

Due to their complex structure, neural networks can represent complicated relationships very well and thus achieve excellent performance. The main advantage over decision trees and especially logit models is that they can approximate nonlinear data structures with a large number of variables with arbitrary accuracy (Huang, 2003). Thus, complexity and intransparency of a data structure can be better revealed (Wiedmann & Jung, 2003, p. 49). At the same time, they have excellent generalization properties and are robust to data imperfections due to their parallelization properties (Zell, 2003, p. 27).

As this paper shows, the advantages of neural networks also manifest themselves in predicting student dropout. To this end, this paper first presents the main results of the initial study in the second chapter. In the third chapter, the data basis is explained. The fourth chapter develops the methodology used. The fifth chapter presents the results. The sixth chapter discusses key findings. The seventh chapter briefly concludes.

2. Main Findings of Initial Study

The initial study by Kemper et al. (2020) investigated the prediction of dropout with decision trees (DT) and logit models (logistic regression, LR) based on study progression data and a few applicant data. The results show that dropout can be identified with these models after the first and after the third semester with an accuracy of 85 and 95percent, respectively (cf. Table 5 in Appendix A). If one considers the sensitivity, which only refers to the correct classification of dropouts, the estimation accuracy drops to 63 and 90 percent. The ROC curves shown in Figure 1 (left) illustrate the improvement of the prediction quality through the addition of further semesters.

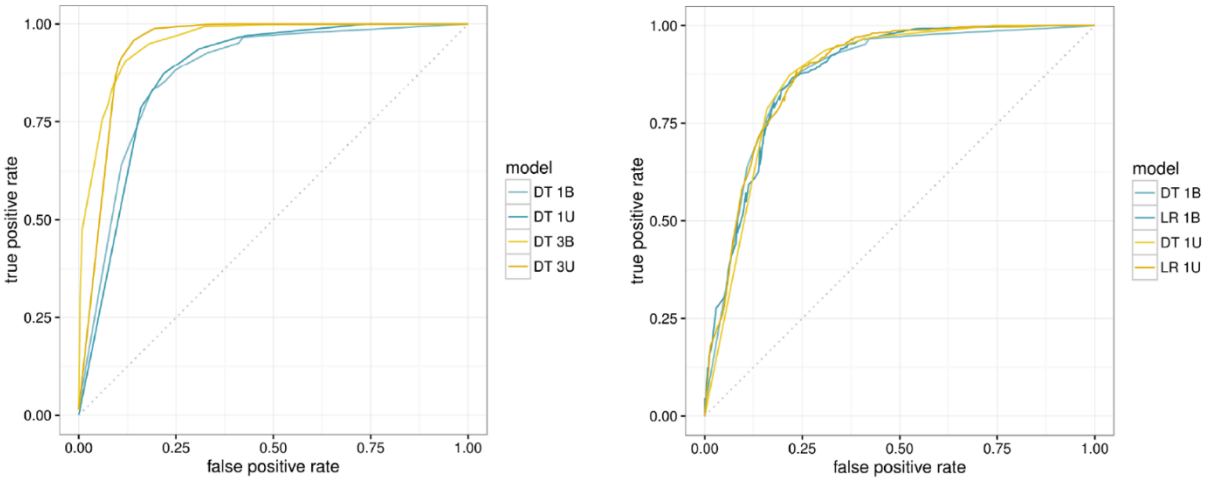

Figure 1: ROC-curves of the decision trees and logit models of the initial study (by Kemper et al. (2020))

The decision trees usually lead to slightly better results in terms of estimation accuracy than the logit models (cf. Table 5 in Appendix A). In relation to the other performance measures, however, no meaningful difference can be discerned. Figure 1 (right) shows exemplarily for the first semester that decision trees and logit models hardly differ.

The dataset used contains significantly fewer dropouts than successes. In order to focus the classification on dropouts, the initial study uses synthetically balanced datasets in addition to the unbalanced datasets. The results are clear: all models trained on balanced datasets classify dropouts significantly better and study successes worse than the models trained on the unbalanced datasets (cf. Figure 1 left).

3. Data

The data basis depends on the initial study. Applicant data considered are gender, age and nationality. The study progress data used are the grade, number of participants, status (passed, failed, failed at last attempt), exam ID and date of the exam for each exam and module. In addition, the enrolment date and final result (completion, dropout, in active study) are known for each student. As in the original study, data is used for the period from October 2002 up to and including October 2016. This results in a data basis of 327,144 observations of examinations and modules of 5,168 students. By 2016, 2,556 students had successfully completed their studies and 620 had dropped out.

Data cleaning, formatting, selection and extraction are identical to the initial study to ensure comparability of results. Table 1 provides an overview of the final variables used for the analysis and their value ranges.

After pre-processing, different datasets were generated for training the neural networks. Table 8 in Appendix B provides a list of the datasets used. The datasets differ in the number of semesters they contain. Since the aim of this work is to make predictions as early as possible but still reliably, three datasets were created for the neural networks, containing only the first semester, the first and second semesters and finally all three semesters.

Table 1: Overview of the feature space

Variable	Type	Value range	Description
Student ID	Nominal		Anonymous ID
Success	Nominal	0 = dropout 1 = success 2 = enrolled	Indicator for student status
Endat	Datum	01.10.07 – 01.01.12	Date of enrolment
Sex_m	Nominal	0 = female 1 = male	Gender
Staat_d	Nominal	0 = not german 1 = german	Origin
Age	Dezimal	17 to 24 per year > 24 per five years	Age at enrolment
Pnr	Nominal		Exam ID
X_pnote	Dezimal	1.0 bis 5.0	Grade in exam X
X_pstatus	Nominal	0 = failed 1 = passed 2 = applied	Result in exam X
X_sem	Nominal	1 bis 3	Semester of exam X
Note_avg	Dezimal	1.0 to 5.0	Average grade in all exams
Avg_be	Dezimal	1.0 to 5.0	Average grade in all passed exams
Sem_max	Nominal	1 to 3	Semester of last exam
P_count	Integer	≥ 0	Number of exams
Be_count	Integer	≥ 0	Number of passed exams
Nb_count	Integer	≥ 0	Number of failed exams

Another aspect in which the datasets differ is the number of variables. A prediction based purely on the extracted variables would offer the possibility of using a much smaller model with a small number of input variables. By limiting the model to aggregate variables, data protection concerns could also be more effectively mitigated in the actual application. In order to evaluate a possible limitation of the model, further datasets were therefore generated, which contain only the extracted variables in addition to the applicant data.

Due to the unequal distribution of dropouts and successful completions (ratio 1:4.13), the initial dataset is unbalanced. In the case of unbalanced datasets, machine methods tend to learn the identification of the majority class (in this case, study successes) preferentially. When predicting dropouts, the aim is to predict students who are at risk of dropping out and therefore, to identify the minority class. Previous studies show that balancing datasets can reduce the classification error related to the minority class and, additionally, increase the accuracy of the classification of the entire dataset (Delen, 2010). Therefore, in order to focus the machine learning procedure on the detection of dropouts, synthetically balanced datasets

were used alongside the unbalanced datasets. Like the initial study, this work uses SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetically balanced datasets (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

In summary, twelve datasets were generated for the neural networks, differing in the number of semesters, in the number of variables and in the use of balanced data.

4. Method

The prediction of dropout is a classification problem with a binary target variable (dropout or success). In order to classify the target variable with neural networks, the applicant and study progress data (raw data) received from KIT were first preprocessed into the final datasets and considered descriptively (cf. chapter Data and Kemper et al. (2020)). A neural network (MLP) was then trained for each of the datasets. This resulted in twelve different models. During training, each dataset was divided into training, validation and test data (70, 20 and 10 percent). For the selection of the hyperparameters, the neural networks were trained on different combinations of values of the hyperparameters with the training data and compared using the validation data. The resulting selected models were then trained using a ten-fold cross-validation with the training and validation data and evaluated using the test data.

4.1. Neural Network Training

Due to their complexity, neural networks are very susceptible to overfitting. To avoid overfitting and to check the robustness of the results on the test data, cross-validation was therefore carried out both in the selection of the hyperparameters and in the training of the resulting networks.

The optimization of the hyperparameters was carried out using random search. For each dataset or neural network, 256 combinations of values were tested. According to the findings of Bergstra & Bengio (2012, pp. 291-293), this number is sufficient to obtain good results compared to the grid search. Each combination of values was evaluated using five-fold cross-validation to reduce variance. Hyperparameters optimized via random search were the learning rate, the number of hidden layers and neurons per layer, the choice of activation function and the amount of dropout per layer. A total of eight to 14 hyperparameters were thus optimized, depending on the number of hidden layers. The other hyperparameters were selected manually. An overview of all hyperparameters and the tested values can be found in Table 9 in Appendix B.

Manual selection

Some hyperparameters could be selected manually. For example, the output layer of all neural networks in this work has a single neuron and a Fermi activation function due to the binary nature of the problem. Furthermore, the neural networks were trained with the backpropagation algorithm. For this purpose, this work used the Adam optimizer by Kingma & Ba (2014). Mean square deviation was used for the loss function.

Another hyperparameter to be selected was the number of epochs over which the neural network training runs. If too few epochs are chosen for training, the neural network will not

achieve the best possible accuracy. In contrast, an unnecessarily high number of training epochs can lead to overfitting of the model (Bengio, 2012, p. 9). For this analysis, therefore, a number of 50 training epochs was used and the principle of early stopping was applied. Here, the training of the neural network is stopped as soon as there is no longer a significant improvement in the generalization error over a previously defined number of epochs. According to Bengio (2012, pp. 9-10), early stopping is not only suitable for avoiding overfitting due to the number of epochs, but also due to the other hyperparameters. This work uses early stopping after ten epochs.

Besides the number of epochs, the size of the batch is an important hyperparameter: A larger batch size allows for a shorter computation time, but also reduces the training progress per epoch, since fewer updates of the weights take place (Bengio, 2012, p. 9). In this work, a batch size of 128 is used.

Selection by Random Search

According to Bengio (2012, S. 8), the learning rate is the most important hyperparameter to be optimized: if it is chosen too large, it misses the minimum of the loss function. It was therefore tested with random values in the interval [0.00001; 0.01]. With a learning rate of 0.01, for example, only one hundredth of the calculated training error per batch or epoch is corrected by adjusting the weights. In addition, to better determine the minimum of the loss function, the principle Reduce Learning Rate on Plateau was used. Here, the learning rate is reduced by a fixed factor as soon as the optimizer could not improve the loss function over a selected number of epochs. In this work, the factor 0.05 and an adjustment after 10 epochs were used.

The number of hidden layers and the number of neurons per hidden layer were also optimized via random search. Figure 2 shows, using the MLP 12B model as an example, that both hyperparameters have a clear influence on the prediction quality of the neural networks. The architecture of the best model in this example consists of two hidden layers and 256 neurons per hidden layer. There is a tendency for a higher number of neurons associated with fewer than five hidden layers to give better results for this model. However, a monotonic relationship between the two hyperparameters is not evident.

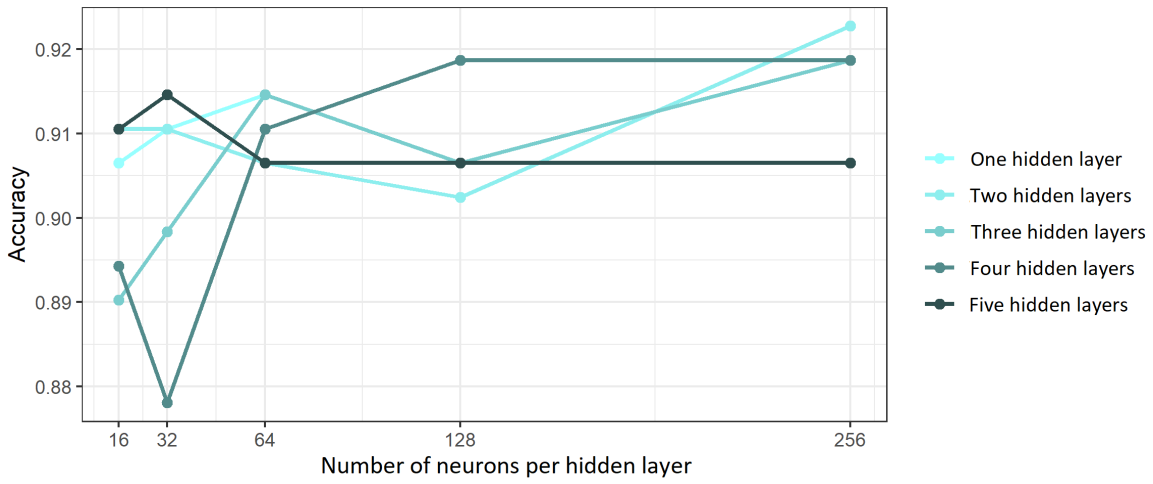

Figure 2: Accuracy of the model 12B with validation data, depending on number of hidden layers and number of neurons per hidden layer

For the choice of the activation function, the functions ReLu, Fermi and TanH were tested.

Finally, the amount of the so-called dropout per layer was optimized via random search. Dropout is a regularization method that prevents the overfitting of neural networks (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). For each computational step, a previously specified number of neurons of the corresponding hidden layer is switched off on a random basis and not considered for the computation. The neurons of the layer thus learn fewer specific concepts and the neural network becomes more robust with respect to unknown data. In this work, a dropout of 0 to 40 percent was tested for each hidden layer.

4.2. Network Topology

This study uses MLP for predicting dropout.[1] MLP are forward-looking networks trained with the backpropagation algorithm. The term "feedforward" is derived from the absence of any feedback within the architecture. Data is read into the network via input neurons, then processed in the hidden layers and subsequently made available to the output neurons without passing through a layer of the network several times or skipping over different layers. Neurons are thus arranged in layers, with each neuron connected to each other neuron of the following layer by a weighted edge.

The optimization of the hyperparameters via random search has resulted in different combinations of values for each of the twelve MLPs. Table 2 contains the resulting network topology of the MLP 123 model. It consists of an input layer, two hidden layers and an output layer. The input layer and the hidden layers are regulated by dropout (see Table 10 in Appendix B for the dropout rates). All four layers have a different number of neurons (64, 160, 32 and 1).

Table 2: Network architecture of the model MLP 123

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	38272
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 160)	10400
dropout_1 (Dropout)	(None, 160)	0
dense_2 (Dense)	(None, 32)	5152
dropout_2 (Dropout)	(None, 32)	0
dense_3 (Dense)	(None, 1)	33

Total params: 53,857
Trainable params: 53,857
Non-trainable params: 0

[1] We also employed recurrent networks to exploit the temporal structure of the data. However, the results showed that sequences of three semesters are too short to achieve a meaningful classification with recurrent networks.

Table 10 in Appendix B includes all individual hyperparameter values of the respective models. The low variation in the learning rate is striking: for most MLPs, a learning rate of 1 percent was found, only the models MLP 12EB, 123B and 123E received 0.1 percent as an optimal value. For the other hyperparameters, however, no tendency towards specific values can be inferred from the results. The best possible number of hidden layers varies in the different models from one to three, with two hidden layers in half of the models. A correlation between datasets with only extracted and all features, between datasets with balanced and unbalanced data or related to the number of semesters considered is not evident. Similarly, the optimal number of neurons per layer varies. The dropout rates vary between 0 and 0.4 and thus completely cover the tested range of values. Finally, no trend is evident for the activation functions either.

5. Results

The neural networks used in this study achieve prediction accuracies of 90.6 percent in the first semester and 98.1 percent in the third semester on the test data. In particular, the identification of dropouts is extraordinarily successful with the corresponding sensitivity values of 94.7 and 95.6 per cent. The corresponding kappa values of 85.0 and 91.9 percent also show an excellent improvement in the predictive quality of the models compared to random models.

In order to exclude a possible overfitting of the models, their training was carried out within the framework of a ten-fold cross-validation. The results show a low variance of the accuracy and loss function in the individual runs (cf. Table 11 in Appendix B). The balanced models tend to show higher deviations than the unbalanced models. However, there is no overfitting of the neural networks in any of the models. Figure 3 shows an example of the results of the cross-validation of the MLP 1 model: the accuracy of the model on the respective test data (dark blue) fluctuates slightly around the value of the accuracy on the training data (light blue).

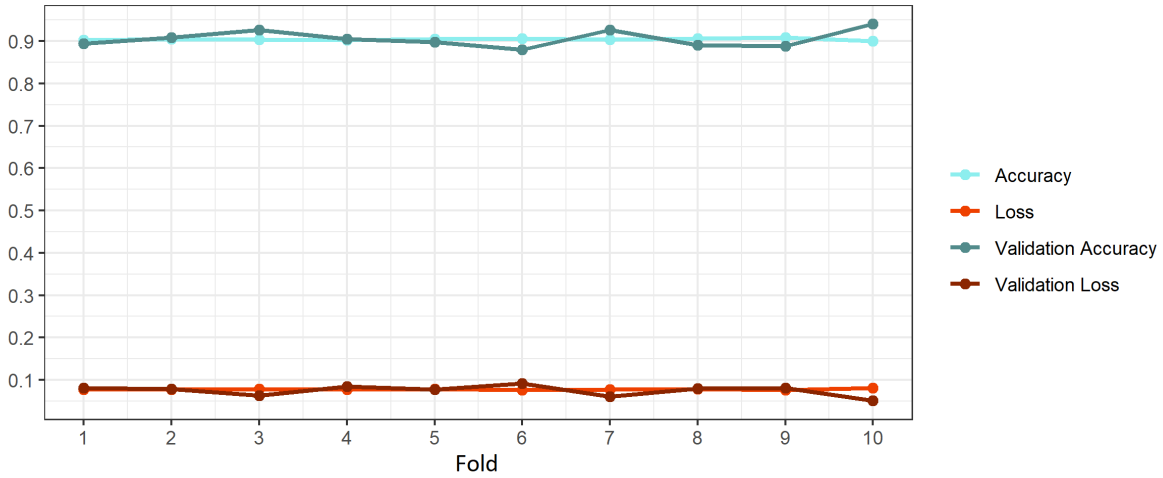

Figure 3: Results of the ten-fold cross validation with test data on model MLP 1

In the following, the results of the different models are considered in detail. Subsequently, the robustness of the models is checked with regard to test data from individual semesters.

5.1. Multilayer Perceptrons

A total of twelve models were generated using MLP and by training on different datasets. The highest accuracy, sensitivity and kappa values were achieved by model MLP 123B, which was trained with synthetically balanced data from the first three semesters including all variables. Thus, this model shows the best identification of dropouts as well as the highest improvement in predictive performance compared to a randomized model. Table 3 contains the complete results of the performance metrics for the individual MLP models. The values are derived from the average of the cross-validation runs.

In all models, the addition of further semesters leads to an improvement in the results. The ROC curves of the models illustrate that with increasing number of semesters significantly better predictions are possible for both study success and dropouts (cf. Figure 4 and Figure 5). Likewise, the values of Accuracy, Sensitivity and Kappa confirm that the overall accuracy, the detection of dropouts and the goodness compared to random models improves strictly monotonically with increasing number of semesters.

Table 3: Results of the performance metrics on all models with test data (highest values marked in blue)

Model	Accuracy	Specificity	Sensitivity	Precision	F1-Score	F2-Score	Kappa	AUC
MLP 1	0.9058	0.9703	0.5532	0.7647	0.6420	0.5856	0.5909	0.8676
MLP 12	0.9380	0.9655	0.7321	0.8200	0.7736	0.7481	0.7283	0.9241
MLP 123	0.9626	0.9656	0.8364	0.8364	0.8364	0.8364	0.8020	0.9511
MLP 1B	0.9063	0.9009	0.9470	0.9191	0.9329	0.9413	0.8503	0.9745
MLP 12B	0.9388	0.9909	0.9044	0.9919	0.9461	0.9206	0.8861	0.9802
MLP 123B	0.9807	0.9640	0.9562	0.9704	0.9632	0.9590	0.9186	0.9850
MLP 1E	0.8875	0.9740	0.5745	0.7941	0.6667	0.6081	0.6191	0.8542
MLP 12E	0.9163	0.9655	0.7500	0.8235	0.7850	0.7636	0.7415	0.9021
MLP 123E	0.9398	0.9656	0.7818	0.8269	0.8037	0.7904	0.7639	0.9375
MLP 1EB	0.8387	0.8468	0.8939	0.8741	0.8839	0.8899	0.7424	0.9288
MLP 12EB	0.8857	0.9273	0.8971	0.9385	0.9173	0.9051	0.8201	0.9612
MLP 123EB	0.9091	0.9369	0.9051	0.9466	0.9254	0.9131	0.8377	0.9832

The MLP trained on unbalanced data and all variables already achieve very good prediction accuracies with accuracy values of 90.6 in the first study semester to 96.3 percent in the third study semester. The identification of study successes is successful with over 96 percent. However, the results for the identification of dropouts are significantly worse. The MLP1 model correctly classifies only 55.3 percent of dropouts. At the very least, the Precision values show that three out of four students classified as dropouts were correctly classified.

With prediction accuracies of up to 98.1 percent, the models trained on balanced and all variables achieve the best results. The high sensitivity values of 94.7 percent in the first study semester and 95.6 percent in the third study semester should be emphasised. Figure 4 shows that synthetic balancing leads to an almost equally accurate identification of dropouts as of study successes. AUC values of over 97 per cent further indicate a nearly perfect assignment of students by the models.

Compared to the models with all variables and unbalanced data, the models with balanced data have consistently better accuracy values. Furthermore, the identification of study discontinuations improves significantly through synthetic balancing: the sensitivity increases by up to 39 percentage points. Likewise, the balanced models with higher AUC values show a higher probability of assigning students to the correct class. In contrast to the non-balanced models, the kappa values are above the limit of 0.8, indicating a very good improvement of the models compared to random models. Figure 4 illustrates that the balanced models outperform the unbalanced models not only in terms of sensitivity. The red ROC curves are consistently below the blue ROC curves. If a cut-off value is chosen so that the sensitivity is higher than 90 percent, the unbalanced models achieve significantly worse specificity values. In summary, the balanced models are clearly better suited for predicting dropout.

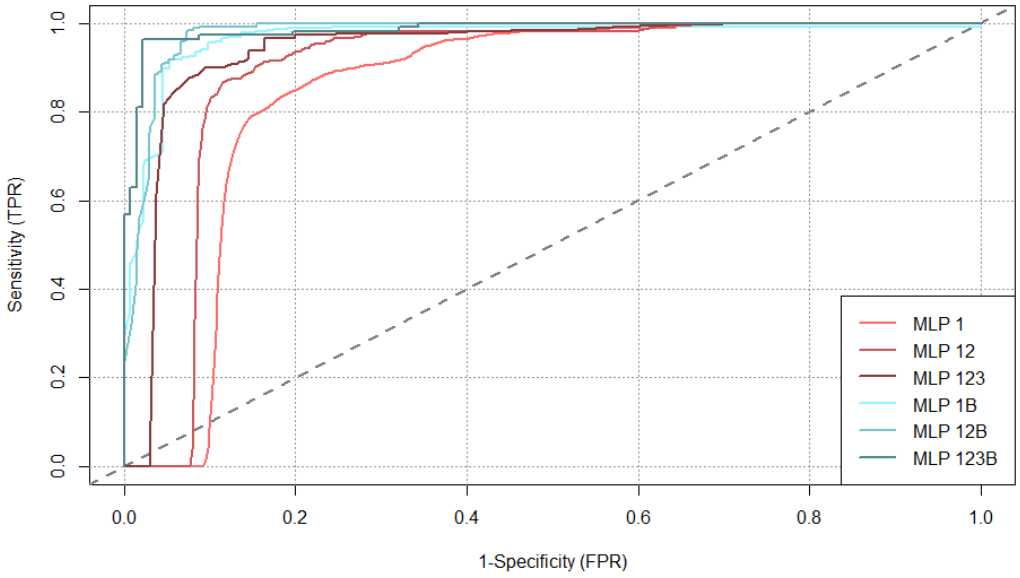

Figure 4: ROC-curves of the models with balanced (blue) and unbalanced (red) datasets with all variables included

The models trained on unbalanced data and exclusively with the extracted variables achieve very good prediction accuracies with accuracy values of 88.7 percent in the first study semester to 93.9 percent in the third study semester. The identification of study successes is successful with over 96 percent. However, the results for the identification of dropouts are significantly worse. The MLP1E model correctly classifies only 57.4 percent of dropouts. The ROC curves in Figure 5 illustrate the strong focus of the models on the correct classification of study successes. At the very least, the Precision values show that students classified as dropouts were correctly classified with a probability of over 79 per cent.

By synthetically balancing the data, the sensitivity values can be increased to 89.3 to 90.1 percent (MLP 1EB, 12EB, 123EB). However, the identification of study success deteriorates by up to 12 percentage points. Figure 5 illustrates the trade-off between increased sensitivity and reduced specificity. The accuracy of the models also decreases. Overall, however, the models improve over random models: the kappa values are significantly higher and are above 0.8 in the second and third semester. Furthermore, the balanced models with higher AUC values show a greater probability of assigning students to the correct class.

The models trained on unbalanced data and exclusively with the extracted variables differ from the models trained on all variables only by a few percentage points. While Accuracy and AUC are slightly worse, Sensitivity and Kappa provide slightly better values, at least for the first two study semesters. Consequently, the prediction of dropout is already successful with little aggregated data. In contrast, the models trained on balanced data and exclusively with the extracted variables show significantly worse results than the models trained on all variables. The sensitivity values increase less and are constantly five percentage points below the values of MLP 1B, 12B and 123B. Furthermore, both the probabilities that a student classified as a dropout was correctly classified (Precision values) and the Accuracy and Specificity values are lower than in all other models. The improvement in the prediction of dropouts achieved by balancing thus deteriorates the models to a disproportionate extent.

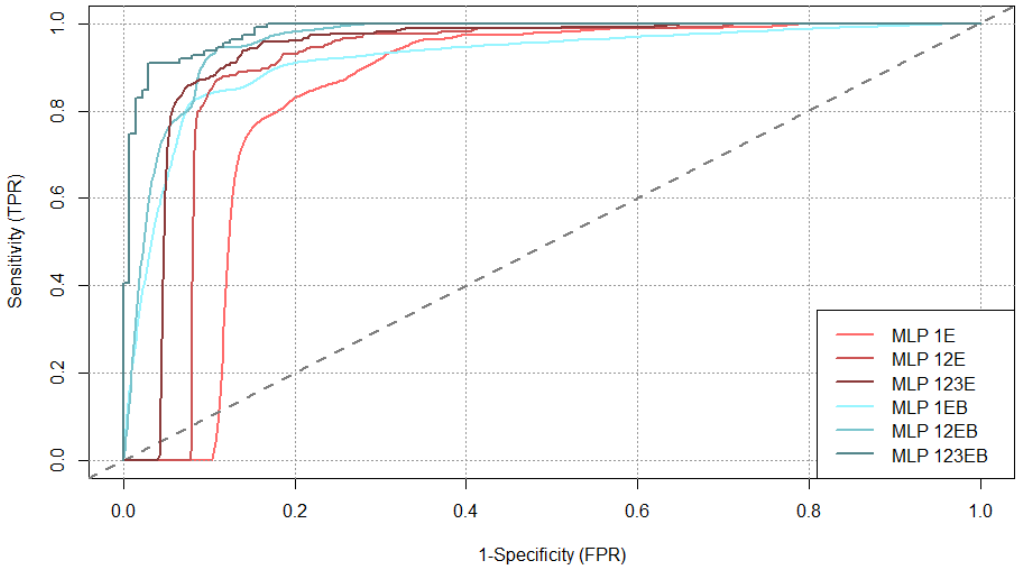

Figure 5: ROC-curves of the models with balanced (blue) and unbalanced (red) datasets with only extracted variables

5.2. Validation per semester

By cross-validation, overfitting of the models could be excluded. The test data was selected on a random basis. For actual use, however, the models must be able to predict student dropout in future semesters based on data from previous semesters. The models must be sufficiently generalised to maintain their predictive quality despite changing conditions within the university (e.g. exam changes). For this reason, this work, like the initial study, undertook a check of the models with training data from previous semesters and test data from the following semester. For example, the MLP 1 2010 model was trained with data up to and including 2009 and validated with 2010 data. Table 4 contains the results of the performance metrics.

The results show that the models are robust to test data from future semesters. In particular, the results with the unbalanced datasets differ little from the results of the MLP 1 model. Compared to the accuracy of 90.6 percent of the MLP 1 model, the accuracy of the models hardly varies. While the specificity of the models is slightly worse than in MLP 1, sensitivity and precision are better in all models than in MLP 1. For the test data from 2009, 2010 and 2012, sensitivity increases by more than 12 percentage points. Similarly, these models provide higher values for kappa. The lower value of kappa and the poorer prediction of dropouts of

the MLP 1 2011 model reflect the fact that there were significantly fewer dropouts among first-year students in 2010 than among the other cohorts. This resulted in a larger share of the majority class in the training dataset.

Table 4: Results of the validation of the models MLP 1 and MLP 1B with test data from previous semesters

Model	Accuracy	Specificity	Sensitivity	Precision	F1-Score	F2-Score	Kappa	AUC
MLP 1 2009	0.8953	0.9374	0.6824	0.6824	0.6824	0.6824	0.6197	0.9140
MLP 1 2010	0.9295	0.9550	0.7612	0.7183	0.7391	0.7522	0.6984	0.9179
MLP 1 2011	0.8991	0.9452	0.5811	0.6056	0.5931	0.5858	0.5356	0.8579
MLP 1 2012	0.8741	0.8943	0.8000	0.6729	0.7310	0.7709	0.6496	0.8869
MLP 1B 2009	0.8588	0.9588	0.7588	0.9485	0.8431	0.7904	0.7176	0.9506
MLP 1B 2010	0.8619	0.9552	0.7687	0.9450	0.8478	0.7985	0.7239	0.9420
MLP 1B 2011	0.7635	0.9662	0.5608	0.9432	0.7034	0.6103	0.5270	0.8796
MLP 1B 2012	0.8722	0.8778	0.8667	0.8764	0.8715	0.8686	0.7444	0.9223

Compared to the unbalanced models, the balanced models are less robust to test data from future semesters. None of the models achieves an accuracy of 90.6 per cent (MLP 1B). Similarly, the values of the other performance metrics are below those of MLP 1B. Only the specificity has better results with over 95 percent in the models MLP 1 2009 to 2011. If, on the other hand, one compares the balanced with the unbalanced models, they still deliver good results. Accuracy is lower, but specificity, sensitivity and kappa tend to have higher values.

6. Discussion

The aim of this work was to improve the prediction quality compared to the models of the initial study by classification with neural networks. This succeeds with both the unbalanced and balanced data sets. While the original study can identify dropouts in the first (third) semester with an accuracy of 76.4 (89.6) percent, the neural networks of this study achieve sensitivity values of 94.7 (95.6) percent. They not only achieve a better identification of dropouts, but also an overall higher prediction accuracy: the accuracy of the neural networks is 90.6 (98.1) percent in the first (third) semester, whereas the best accuracy value of the initial study is 88.1 (95.3) percent (cf. Table 3 and Table 5 in Appendix A). Furthermore, the kappa values show that the neural networks already achieve a very good improvement over a random model from the first semester onwards. With the decision trees and logit models, on the other hand, only the kappa values of the models LR 3U, DT 3U and DT 2B are above the limit of 0.8.

The different results for the models with synthetically balanced and unbalanced data should be emphasised. The accuracy of the neural networks is higher in all models than that of the decision trees and logit models (cf. Table 3 and Table 5 in Appendix A). In addition, the accuracy of the balanced models of this work increases compared to the accuracy of the unbalanced models, while the accuracy of the models of the initial study decreases. In contrast, sensitivity and precision of the neural networks using unbalanced data are lower than for decision trees and logit models. Similarly, kappa assumes worse values for neural networks. For unbalanced data, decision trees and logit models therefore deliver better results in terms of identifying study dropouts. Neural networks, on the other hand, offer

significantly better predictions overall for balanced models. The improvement of the prediction quality through synthetic balancing of the data is consistent with the findings of (Delen, 2010).

To ensure the robustness of the decision trees with respect to the training and test data, the initial study (as well as this paper) performed a ten-fold cross-validation. The results of both studies demonstrate the robustness of the models. Nevertheless, there is a higher difference between the accuracy values of the decision trees than between those of the neural networks (cf. Table 6 in Appendix A and Table 11 in Appendix B). The model DT 1U, for example, shows a deviation of 9 percentage points between minimum and maximum accuracy, while the deviation for model MLP 1 is only 6 percentage points. Consequently, neural networks are better suited to generalise the problem.

Finally, both the original study and the present study validated the models with training data from previous semesters and test data from the following semester. While the accuracy values differ only slightly, the identification of study dropouts is significantly better with the neural networks (cf. Table 4 and Table 7 in Appendix A). The sensitivity is 9 (test semester 2009) to 42 (test semester 2012) percentage points higher than with the decision trees. Furthermore, the consistently higher kappa values of the neural networks illustrate that the models of this work still perform well in this test scenario compared to random models, whereas the models of the initial study only achieve a moderate improvement.

In summary, the neural networks of this work not only offer more accurate predictions than the models of the original study, but also better generalisation properties and more robust results.

7. Conclusion

In the context of this work, prediction models for dropout were developed with the help of neural networks and study progress data. These models enable predictions regarding study success and dropout with an accuracy of up to 91 and 98 percent in the first and third semester, respectively. In particular, the identification of dropouts is already successful in the first semester with a sensitivity of 95 percent. Compared to the initial study, the prediction of dropouts is thus improved. This confirms the thesis that neural networks are better able to grasp the complexity of the problem and at the same time achieve excellent generalisation.

Compared to the methods of the initial study, the neural networks show better forecasting accuracies. Nevertheless, the decision trees and logit models have non-negligible advantages over neural networks. For example, neural networks are significantly more difficult to analyse due to their complexity and, in contrast to decision trees and logit models, they do not offer any explanatory components, so that their decisions are not comprehensible. However, especially when predicting dropout, an explanation of the decision is important in order to identify possible causes or to be able to advise students in a meaningful way. With regard to the actual use of the models in higher education, a combination of these machine methods is therefore recommended.

References

Bengio, Y. (2012). Practical Recommendations for Gradient-Based Training of Deep Architectures. In *Neural networks: Tricks of the trade* (pp. 437-478). Springer, Berlin, Heidelberg.

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk--Predicting Student Dropouts Using Administrative Student Data form German Universities and Machine Learning Methods. *Journal of Educational Data Mining, 11(3)*, pp. 1-41.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research, 13(1)*, 281-305.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, pp. 321-357.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49(4)*, 498-506.

Hetze, P. (2011). *Nachhaltige Hochschulstrategien für mehr MINT-Absolventen.* technical report, Stifterverband für die Deutsche Wirtschaft; HeinzNixdorf-Stiftung.

Heublein, U. (2014). Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen: Statistische Berechnungen auf der Basis des Absolventenjahrgangs 2012. *Deutsches Zentrum für Hochschul- und Wissenschaftsforschung*.

Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J., & Woisch, A. (2017, Juni). *Zwischen Studienerwartungen und Studienwirklichkeit.* Hannover: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung GmbH.

Huang, G. B. (2003). Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Transactions on Neural Networks, 14(2)*, pp. 274-281.

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting Student Dropout: A machine learning approach. *European Journal of Higher Education, 10(1)*, 28-47.

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980.

Ram, S., Wang, Y., Currim, S. A., & Currim, F. A. (2018). *Predicting student retention using smartcard transactions.* United States Patent Application Publication. US Patent 2018/0144352 A1.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout. A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research, 15(2)*, 1929-1958.

Wiedmann, K.-P., & Jung, H. H. (2003). *Neuronale Netze im Marketing-Management - Eine praxisorientierte Einführung in modernes Data-Mining.* (K. P. Wiedmann, Ed.) Wiesbaden.

Zell, A. (2003). *Simulation neuronaler Netze.* 4. unveränderter Nachdruck, Bonn.

Appendix A

Table 5: Results of the initial study for all models (Kemper, Vorhoff, & Wigger, 2020)

	Model	ACC	SEN	SPC	PRE	KAP	TP	TN	FP	FN
Logistic Regression	LR 1U	0.881	0.629	0.960	0.828	0.642	173	853	36	102
	LR 2U	0.920	0.751	0.974	0.903	0.769	214	871	23	71
	LR 3U	0.945	0.785	0.984	0.924	0.816	256	1320	21	70
	LR 1B	0.845	0.775	0.866	0.642	0.598	213	770	119	62
	LR 2B	0.896	0.839	0.914	0.756	0.726	239	817	77	46
	LR 3B	0.918	0.862	0.932	0.755	0.754	281	1250	91	45
Decision Tree	DT 1U	0.879	0.691	0.937	0.772	0.652	190	833	56	85
	DT 2U	0.930	0.765	0.982	0.932	0.796	218	878	16	67
	DT 3U	0.953	0.807	0.988	0.943	0.841	263	1325	16	63
	DT 1B	0.865	0.764	0.897	0.695	0.639	210	797	92	65
	DT 2B	0.933	0.828	0.966	0.887	0.813	236	864	30	49
	DT 3B	0.918	0.896	0.924	0.741	0.760	292	1239	102	34

Table 6: Accuracy of the decision trees of the initial study (not balanced models) on test data with ten-fold cross-validation (Kemper, Vorhoff, & Wigger, 2020)

model	fold-1	fold-2	fold-3	fold-4	fold-5	fold-6	fold-7	fold-8	fold-9	fold-10
DT 1U	0.888	0.897	0.846	0.905	0.846	0.862	0.862	0.812	0.845	0.855
DT 2U	0.940	0.932	0.873	0.890	0.932	0.924	0.958	0.898	0.924	0.932
DT 3U	0.940	0.928	0.934	0.928	0.898	0.940	0.964	0.922	0.892	0.910

Table 7: Results of the validation of the model DT 1U with test data from previous semesters (Kemper, Vorhoff, & Wigger, 2020)

Semester	ACC	SEN	SPC	PRE	KAP
10.01.2009	0.893	0.588	0.954	0.714	0.583
10.01.2010	0.908	0.433	0.980	0.763	0.505
10.01.2011	0.897	0.257	0.990	0.792	0.347
10.01.2012	0.862	0.378	0.994	0.944	0.476

Appendix B

Table 8: Overview of all datasets used for the classification

Dataset	Included semester	Included variables	Balanced data
1	Semester 1	All variables	No
1B	Semester 1	All variables	Yes
1E	Semester 1	Extracted variables	No
1EB	Semester 1	Extracted variables	Yes
12	Semester 1 to 2	All variables	No
12B	Semester 1 to 2	All variables	Yes
12E	Semester 1 to 2	Extracted variables	No
12EB	Semester 1 to 2	Extracted variables	Yes
123	Semester 1 to 3	All variables	No
123B	Semester 1 to 3	All variables	Yes
123E	Semester 1 to 3	Extracted variables	No
123EB	Semester 1 to 3	Extracted variables	Yes

Table 9: Type of selection and value range for the hyperparameter tuning

Hyperparameter	Selection type	Value range
Number of hidden layers	Random search	1 to 4
Number of neurons per layer	Random search	16 to 256
Optimization function	Manuel	Adam, SGD, RMSprob
Activation function	Random search	Fermi, ReLu, Tanh
Loss function	Manuel	Mean squared error, binary crossentropy
Learning rate	Random search	0.1 to 0.00001
Dropout	Random search	0 to 0.4
Batch	Manuel	32, 64, 128
Number of training epochs	Manuel	20, 50, 100

Table 10: Resulting hyperparameter values of all MLP models after applying random research and achieved accuracy on validation data

Model	**ε**	***h***	**Number of neurons**	**Activation function**	**Dropout**	***Accuracy***
MLP 1	0.01	2	32/96/64	tanh/relu/relu	0.35/0.15/0.35	0.9082
MLP 1B	0.01	3	32/32/192/64	tanh/tanh/relu/fermi	0.1/0.1/0/0.35	0.9259
MLP 1E	0.01	3	32/192/96/64	tanh/tanh/tanh/fermi	0.15/0.1/0/0.1	0.9146
MLP 1EB	0.01	2	128/128/32	tanh/tanh/relu	0.1/0.05/0.2	0.8724
MLP 12	0.01	2	32/32/224	relu/relu/relu	0.2/0.3/0.25	0.9243
MLP 12B	0.01	1	64/96	tanh/tanh	0.3/0.4	0.9431
MLP 12E	0.01	2	128/96/128	tanh/tanh/fermi	0.3/0.15/0.3	0.9274
MLP 12EB	0.001	2	224/160/160	fermi/tanh/tanh	0.15/0.25/0.15	0.9106
MLP 123	0.01	2	64/160/32	relu/tanh/relu	0.15/0.05/0.15	0.9432
MLP 123B	0.001	1	160/160	tanh/tanh	0.05/0.3	0.9597
MLP 123E	0.001	3	128/96/64/64	relu/tanh/relu/relu	0.15/0/0.15/0.35	0.9338
MLP 123EB	0.01	1	96/192	fermi/relu	0.35/0.4	0.9194

Table 11: Accuracy of the MLP models on test data with ten-fold cross-validation

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
MLP 1	0.8944	0.9085	0.9263	0.9049	0.8982	0.8803	0.9263	0.8877	0.9401	0.9058
MLP 1B	0.9132	0.9041	0.9087	0.8991	0.8767	0.8904	0.9406	0.9269	0.8904	0.9132
MLP 1E	0.8592	0.8838	0.8912	0.8803	0.8912	0.8697	0.9088	0.8979	0.8772	0.9155
MLP 1EB	0.8676	0.8676	0.8037	0.8532	0.8128	0.8311	0.8447	0.8265	0.8311	0.8493
MLP 12	0.9266	0.9371	0.9544	0.9301	0.9263	0.9298	0.9509	0.9298	0.9404	0.9545
MLP 12B	0.9776	0.9324	0.9054	0.9459	0.9595	0.9058	0.9414	0.9324	0.9685	0.9189
MLP 12E	0.9161	0.9161	0.9228	0.9056	0.9053	0.9193	0.9088	0.9123	0.9298	0.9266
MLP 12EB	0.9148	0.8694	0.8604	0.9144	0.8739	0.8610	0.8739	0.8919	0.9324	0.8649
MLP 123	0.9545	0.9720	0.9580	0.9510	0.9825	0.9615	0.9615	0.9720	0.9474	0.9650
MLP 123B	0.9777	0.9910	0.9686	0.9910	0.9865	0.9821	0.9596	0.9955	0.9821	0.9731
MLP 123E	0.9371	0.9545	0.9476	0.9301	0.9580	0.9371	0.9301	0.9510	0.9193	0.9336
MLP 123EB	0.8705	0.9372	0.8879	0.9238	0.9103	0.9238	0.9058	0.9152	0.9283	0.8879