Education 4.0 – Fostering Student's Performance with Machine Learning Methods

Monica Ciolacu[1,2], Ali Fallah Tehrani[2], Rick Beer[2] and Heribert Popp[2]

[1] UPB CETTI, University Politehnica of Bucharest, Romania
[2] Faculty of Business Informatics, Deggendorf Institute of Technology DIT, Bavaria, Germany
monica.ciolacu@th-deg.de

***Abstract*—Educational activity is increasingly moving online and course contents are becoming available in digital format. This enables data collection and the use of data for analyzing learning process. For the 4th Revolution in Education, an active and interactive presence of students contributes to a higher learning quality. Machine Learning techniques recently have shown impressive development steps of the use of data analysis and predictions. However, it has been far less used for assessing the learning quality. For this paper we conducted analysis based on neural networks, support vector machine, decision trees and cluster analysis to estimate student's performance at examination and shape the next generation's talent for Industry 4.0 skills.**

Keywords—***Education 4.0; learning analytics; machine learning; neural networks; k-means clustering; kernel methods.***

I. INTRODUCTION

The great potential of Industry 4.0 lies in data and in the efficient use of newly gained opportunities and challenges. Three trends [1] appear in the Gartner "Hype Cycle for Emerging Technologies 2017": Artificial Intelligence (AI) Everywhere, Transparent Immersive Experiences and Digital Platforms. Some results of this report are presented in Fig. 1.

The impact of digital transformation on industries will displace many tasks and activities which have been traditionally performed by human beings [2]. Taking advantage of existing data needs a change in mindset [3,4].

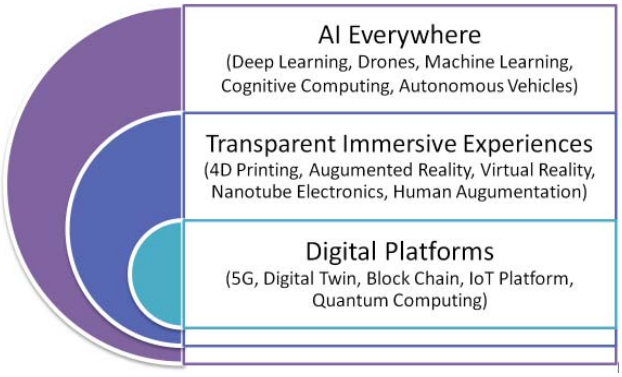

Fig.1. The three trends that will drive the digital economy [1]

As digitalization disrupts the workforce, we consider a large demand for qualified employees for Industry 4.0 including managing knowledge obsolescence and achievement gaps [4].

The demand for a revolution in education becomes stronger. Researches in AI have recently shown impressive leaps in development, with the result that machine learning introduced many AI-based techniques in the industry [5]. AI is shaping the future of everything from medicine, to transportation, to manufacturing and to education [6]. Machine learning is a part of AI that offers the ability to extract new knowledge and patterns from data with huge benefit potential (see Table I).

TABLE I. EVOLUTION OF NEURAL NETWORKS AND BI [1]

Neural networks and Modern Business Intelligence (BI) Platforms Evolution with Data and Analytics	**year**
Companies invested 26.mil $ to 39.mil $ in AI	2016
Deep learning jobs grew from 0 to 41.000 jobs since 2014	2017
Deep Learning (deep neural networks) will be a standard component in 80% of data scientists' toll boxes	2018
Natural Language Generation will be a standard feature of 90% of modern BI and analytics platforms	2019

Artificial Intelligence will play a key role in Education identifying new drivers of students' performance and early disengagement cues, adopting personalizing learning, answering students' routine questions, using learning analytics and providing predictive modeling.

II. DEFINITION OF EDUCATION 4.0

We characterize Education 4.0 by virtual courses including an interactive presence in the form of Blended Learning and seven AI driven features as significant challenges [4] in the educational technology: personalized learning process, game based learning using Virtual Reality/Augmented Reality (VR/AR), communities of practice, adaptive technologies [7], learning analytics, intelligent Chabot's and E-Assessment as you can see in Fig.2.

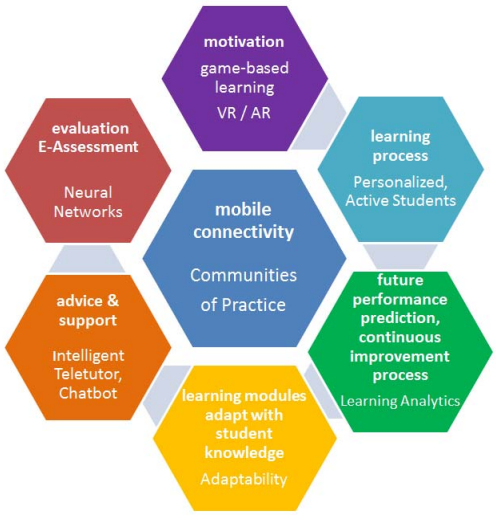

Fig. 2. Technological elements & tools of Education 4.0 with didactic goals

Our research group is currently conducting experiments considering all these features of future virtual courses of the generation Education 4.0 [4]. Particularly, in this paper we focus on Learning Analytics (LA).

There is an obvious relationship between participation rate and passing the examination. In 2013 Deggendorf Institute of Technology (DIT) began to track its students of STEM courses. We have been monitoring the participation rate for each student individually. Learning analytics proved effective in motivating and engaging students to improve their skills and it fosters academic success. R. Ferguson has developed a vision of the future of LA in her recent paper [8].

III. Data Preprocessing

The pioneer in the field of computer gaming and AI, Arthur Samuel gives the first definition of Machine Learning as the field of study that gives computers the ability to learn without being explicitly programmed [9].

A. Extraction of Logfiles from Learning Management System

For the subsequent analysis of students learning behavior and its correlation with success in exams, we use log data provided by the popular Learning Management System (LMS) Moodle, in which the courses are hosted and operated during the semester. The system creates a log entry for each interaction of a student with the course, e.g. accessing the main page, opening resources or reading discussion threads. The information is recorded in a database table with additional attributes such as the requested material's internal category and the date and time of the access. In Table II an exemplary log entry is illustrated:

TABLE II: Exemplary Log Entry from Moodle

id	**4678869**
username	ab12345
timecreated	1456821071
action	viewed
target	course_module

The first step is to extract the log data for a specific course and export it into a .csv file for further preprocessing the data. Due to data privacy regulations, all data which is extracted must be anonymized by using a hashing algorithm like MD5 [10] on the attribute "username". We use Microsoft Excel in the following steps for transforming the raw .csv data into the needed format for the subsequent tasks. The "timecreated" attribute is a UNIX timestamp and can be easily converted to a human readable format.

B. Data Aggregation and Feature Generation

For the analysis, we need to aggregate the data for each user and generate input features which represent the user's activity in a specific period of time. For neural networks and cluster, we achieve this by using a pivot table and accumulating all log entries for each user in each month of the semester. The resulting data looks like the following sample (Table III) taken from a mathematic course during the winter semester (October – January):

TABLE III: Exemplary data point after aggregation

Username (MD5 hashed)	**Oct.**	**Nov.**	**Dec.**	**Jan.**
354fef4daa4dfa4	102	20	115	190

The four features for each user represent the number of clicks in the course (i.e. any type of logged interaction within the course) in the corresponding month. Thus we have generated numbers which indicate the activity level during the semester and can be used as input features in the learning tasks.

C. Blended Learning Course

The first type of course is Blended Learning, which refers to a mixture of traditional teaching in classrooms with E-Learning elements. We use log data from the Mathematics Course of Business Students at DIT in the winter semester 2015/16 for learning and evaluate our model with the data from the same course in winter semester 2016/17.

We have a total of 115 data points for training (in the format as described in Section III) as input data and the corresponding exam result as a binary output variables (exam passed or failed). Meanwhile it is worth mentioning that we only consider those students, who took the exam, like the actual grade or information about the (non-) participation. Students who did not register for the exam are difficult to distinguish from other course participants (e.g. tutors and technical administrators) in the log data, so we decided to leave them out and avoid additional noise in the data.

D. Complete Virtual Course

For the second type of course, we analyze the data from a Complete Virtual Course, where all activity is occurring online

without considering traditional classroom. The data is taken from the Math Course of Amberg-Weiden University (AWU) during the summer semester 2016 (as training set) and summer semester 2017 (for evaluation). The number of data points for training is 67, the output variable is defined as exam passed / not passed as in Section C.

IV. Binary Classification Methods

A. Problem Setting

As previously mentioned the main objective of this paper is to improve the quality of learning, specifically by tracking students. In particular, we are interested in identifying weakness during the learning process, especially in terms of students-weakness. In essence, this problem is mathematically trackable with Machine learning methods (Fig. 3).

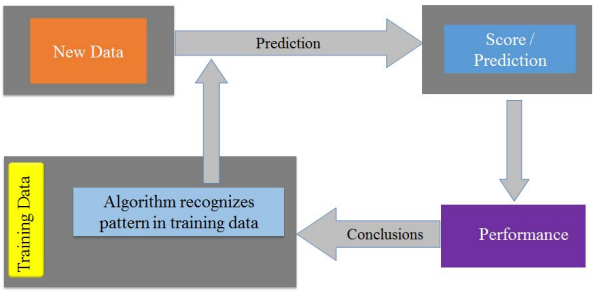

Fig. 3. Machine learning process

More concretely, we apply binary classification. In the following we describe the methods, which we employ for our problem [11]. In the case of binary classification, every instance is labeled by a label from {-1,+1}. More formally, the training data is given as follows:

$$D = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{n} \subset X \times \{-1, +1\} \quad (1)$$

in which D is supposed to be independent and identically distributed, the task is to induce a classifier $L{:}X \rightarrow \{-1,+1\}$, which minimizes the corresponding risk function. For our case, class -1 is assigned to fail the examination and class +1 is assigned to pass the examination.

For the binary classification goal we applied four different approaches, namely, neural networks, kernel methods (support vector machines), decision trees and cluster analysis. In the following we give an overview of these methods:

B. Neural networks

We use neural networks [12] to learn from the existing log data and predict exam results based on the learning activity (Fig. 3).

C. Kernel Methods (SVM)

It is apparent that, if the instances are not linearly separable, then linear support vector machine cannot solve the problem properly, i.e., some instances are miss-classified. In order to prevent miss-classification [11], the basic idea is to transfer the data (instances) to an upper space, of course with higher dimensionality, where the labeled instances can be separated linearly without any mistake. To this end, the core idea is to use kernels which can model the non-linear decision boundaries. Given n training examples the goal underlying kernel machine in dual form is to determine $\{\alpha_i\}$ as follows [13]:

$$\boldsymbol{\alpha} \leftarrow \min\left\{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) - \sum_{i=1}^{n} \alpha_i\right\} \quad (2)$$

Such that:

$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

$$0 \le \alpha_i \le C, \forall i \in \{1, \dots, n\}\,. \quad (3)$$

For the experiments we applied the Polynomial kernel:

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + \gamma)^d \quad (4),$$

where d, the degree of polynomial kernel is a natural number and moreover $\gamma \in R$. In addition the radial basis function (RBF-kernel) is used:

$$k_\sigma(\boldsymbol{x}, \boldsymbol{y}) = exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{2\sigma^2}\right) \quad (5).$$

D. Decision Trees

Now we introduce a top-down greedy approach called Iterative Dichotomiser 3 (ID3) by J. R. Quinlan [14, 15]. Roughly speaking, any decision tree attempts to partition the input space into several proper sub-input spaces, which elements in each sub-input space have a similar output-characteristic. During partitioning decision tree conducts a model on each sub-input space and, furthermore, the algorithm develops incrementally binary trees by adding two edges into each node. As it will be clarified, through partitioning, indeed, decision trees build non-linear decision boundaries, and henceforth, it can model as well as attribute-interaction. Before introducing the algorithm, we shall introduce the concept of entropy. From a statistical point of view, uncertainty quantification in data (S) is calculated by Entropy:

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (6),$$

where X is the set of classes in S and $p(x)$ determines the proportion of the number of elements in class x to the number of elements in S. In the case of ID3, a decision tree is composed by several nodes starting from root node. Given a set of attributes the root node is assigned to an attribute which has the smallest entropy (largest information gain). Once the root node has been identified, the algorithm extends the tree by adding the next attribute which has the smallest entropy among of rest attributes. The algorithm continues to recourse on attributes which never met so far. The algorithm is terminated then on each node if either every element in the subset belongs to the same class, or there are no more attributes or the node does not meet any training example.

A decision tree can be expressed as several rules, for instance:

$$\text{IF } x_j > S \text{ THEN output } = C_1,$$
$$\text{IF } x_j \leq S \text{ AND } x_k > T \text{ THEN output } = C_2.$$
$$\text{IF } x_j \leq S \text{ AND } x_k \leq T \text{ THEN output } = \ldots. \quad (7)$$

E. Cluster Analysis

The final approach is Cluster Analysis, an unsupervised learning technique for identifying structures within datasets and grouping data points based on their characteristics. More precisely, the aim is to create groups of elements "Clusters" so that the similarity within each group is as high as possible, whereas the similarity between the groups should be as low as possible [16]. In our context, we try to identify different groups of students who share a similar learning behavior and determine the average exam result for each group based on the training data. For predicting exam results, we calculate the closest Cluster for a new data sample and interpret the average result of this group as forecast value.

V. Experimental Part SVM and Decision Trees

We conducted several experiments concerning real datasets derived from the course evaluation. Our objectives have been: making a forecast for a new semester and analyzing contents which have been used frequently.

A. Settings

Before applying any machine learning technique, the datasets have been analyzed. To this end, we measure the collinearity between features by using correlation measure. In the correlation values in terms of Pearson for summer semester 2016 (Fig. 4) and summer semester 2017 (Fig. 5) for AWU are depicted. As it can be seen, some of attributes are correlating highly. Technically, collinearity may cause unstable solutions during learning process. From this perspective, redundant features have to be filtered out. Legend: ATT= attempt (quiz), CH= chapter, T= total, C= course main page.

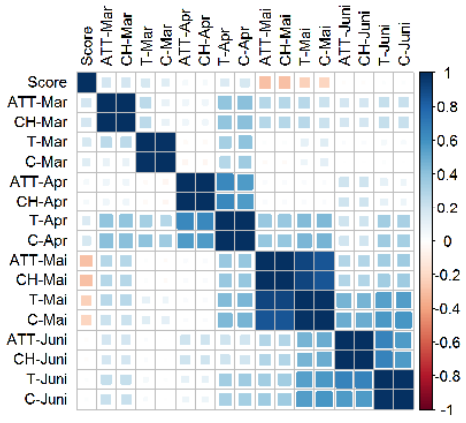

Fig. 4. Pearson Correlation values summer 2016 AWU

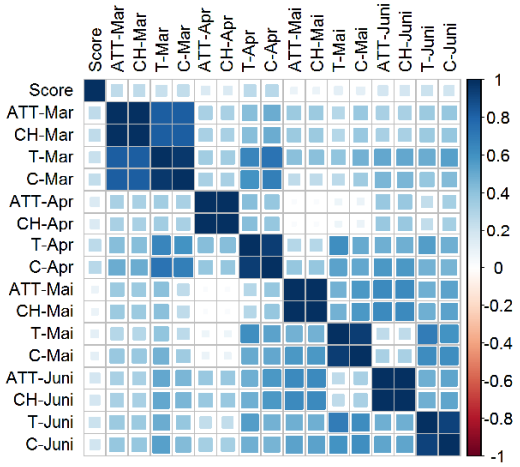

Fig. 5. Pearson Correlation values summer 2017 AWU

For the experiments the following attributes have been taken into account:

- AWU in SS 2017: Total March, Total May, Total June, Total July, Total Results, Course Module, Course March, Course May and Course June.
- DIT in WS 2017: Total October, Total November, Total December, Total January, Total Results, Course October, Course November, Course December and Course January.

The two classes, namely passed or failed are derived from threshold on scores; if it is larger than 40 in a range from 0 to 90, students passed examination otherwise failed. Now instances are incorporating with two classes, and therefore, any binary classification approach can be used. For this study, we apply polynomial kernel, RBF-kernel and finally decision trees. The C-parameter, namely trade-off parameter for kernel machines has been chosen among $\{10^{-3},\ldots,10^{3}\}$ with factors of 10. The sigma parameter for Gaussian kernel has been chosen from $\{10^{-4},\ldots,10^{0}\}$. To verify the quality of prediction we employed accuracy measure as follows:

$$ACC = \frac{\#TP+\#TN}{\#TP+\#TN+\#FP+\#FN} \quad (8),$$

where TP, TN, FP, FN are true positive, true negative, false positive and false negative respectively.

B. Results

An overview of results is illustrated in Table IV. The results are presented the accuracy of each model. For the AWU the Polynomial of degree 3 is superior; this confirms that the dataset has a non-linear characteristic. On contrary, apparently the DIT demonstrates less non-linear characteristic and hence the linear model produces the best results. This may be due to the fact that basically these two universities are using two distinguished models for teaching the students.

TABLE IV: Results of RBF, SVM and Trees

methods	Accuracy Measure	
	AWU SS 17	*DIT WS 17*
Poly d=1	67%	77,3%
Poly d=2	67%	76%
Poly d=3	82,6%	76%
RBF	67%	77,3%
Trees	61,5%	69%

VI. Experimental Part Neural networks

In the following section, we use neural networks provided by the Python library Keras [17] to learn from the existing log data and predict exam results based on the learning activity. We analyze two different scenarios and present the results and conclusions for each course type.

A. Blended Learning Courses

The first type of course is Blended Learning from DIT. The neural network topology providing the best results has the following settings:

- four input neurons (illustrating the number of clicks in Oct., Nov., Dec., Jan.)
- hidden layer with four neurons (activation function: Sigmoid)
- two output neurons with the output value between 0 and 1 (activation function: Softmax) indicating the probability of a data point belonging to one of the output classes (exam passed or not passed)

The number of training epochs is 200, the batch size (number of samples per iteration) is 10. We use the optimizer ADAM [17] for backpropagation learning and Categorical Cross-Entropy as Loss function.

After training the neural network with the data from the winter semester course 2015/16, we use it to predict the exam results for the course in winter semester 2016/17, which has the same course structure. Since we also have the actual exam results from that course, it is possible to determine the prediction accuracy of the neural network. Furthermore we compare the accuracy of the model as described above with two slightly modified versions:

- Using just two input neurons for the first two months and ignoring the rest of the semester ("Mod. 1")
- Using two input neurons per month (4x2=8 in total) by splitting the total number of clicks into the clicks on exercises and all other clicks in the course (see Table V Model 2)

The motivation behind these modified scenarios is to figure out whether the beginning of a semester contains more significant information than the rest, and if the activity in online exercises, which are considered an important preparation for the exam, is more important than the other activities.

The results and the comparison of prediction accuracy are presented in Table V:

TABLE V: Prediction accuracy for Blended Learning Course

Main Model:	4 Month	4 Inputs	Accuracy **70.1%**
Model 1:	2 Month	2 Inputs	Accuracy **75.0%**
Model 2:	4 Month	8 Inputs	Accuracy **61.3%**

We have an accuracy of ~70% in the Main Model and an even better performance of 75% in the modified version with just the first two months as inputs. It seems that the student activity in the beginning of the semester is more significant for the later exam result than the other months. The modification with separate treatment of activity in online exercises however provides worse performance, which might be caused by the course format Blended learning, where a lot of exercises are done outside of the E-Learning course.

B. Complete Virtual Courses

In this section, we analyze the data from the Complete Virtual Course from AWU from the summer semester 2016 (as training set) and summer semester 2017 (for evaluation). The number of data points for training is 67, the output variable is defined as exam passed / not passed as in Section A.

The neural network topology providing the best results has the following settings:

- five input neurons (representing the number of clicks in Mar., Apr., May, Jun., Jul.)
- hidden layer with six neurons (activation function: Sigmoid)
- two output neurons (activation function: Softmax) indicating the probability of a data point belonging to one of the output classes (exam passed or not passed)

The parameters and training functions are chosen according to Section A. Furthermore, we use the same procedure for determining the prediction accuracy and compare the results from the Main Model with the two modifications as described before.

The results for the Complete Virtual Couse are as Table VI:

TABLE VI: Prediction accuracy for Complete Virtual Course

Main Model:	5 Month	5 Inputs	**Accuracy 75.2%**
Model 1:	2 Month	2 Inputs	**Accuracy 64.9%**
Model 2:	5 Month	10 Inputs	**Accuracy 76.2%**

The overall accuracy of 75.2% is slightly higher for the Complete Virtual Course compared to Blended Learning. The modification where only the first two months are considered offers the worst accuracy, namely 10% less than others. It

appears that students in Complete Virtual Courses have a more constant learning behavior and the beginning of semester is not as significant for Blended Learning. The other modification with separate treatment of exercise activity has about the same performance as our "Main Model", so we conclude that there is no benefit in separating the exercise clicks from the overall number of clicks.

VII. Experimental Part Cluster Analysis

In this section, we present the results of the prediction based on Cluster Analysis for both course types Blended Learning and Complete Virtual Course. For Clustering, we use the X-Means-Algorithm, a variation of the popular K-Means-Algorithm which automatically finds the number of clusters in the training data [19]. Each data point consists of two attributes: the overall number of clicks in the course and the number of clicks on virtual exercises. These features have shown good performance in previous analysis [20].

For Blended Learning, the data of DIT in 2015/16 is used. The application of X-Means results in a division into two clusters. For Complete Virtual Courses, we analyze data from AWU 2016. Here the resulting division is three clusters. For each cluster, we also calculate the average exam result of the students belonging to it. To evaluate the prediction accuracy, we assign each student in the corresponding test sets (DIT 2016/17 and AWU 2017) the exam result of the closest Cluster and compare it to the actual result. The prediction accuracy values are presented in Table VII:

TABLE VII: Prediction accuracy for Cluster Analysis

Course Type	Accuracy Measure
Blended Learning	66.3%
Complete Virtual	67.7%

VIII. Conclusion and Outlook

Both types of experiments and analysis have shown that students who begin to learn earlier in the semester have a higher chance to pass the exam. The prediction accuracy is higher for Complete Virtual Course than for Blended Learning. If we compare the evaluated methods we find that non-linear kernel methods and neural networks are superior in terms of prediction accuracy.

Our goal is to identify students at risk early in the semester and to personally motivate them from teacher with mails to enhance their academic success. Another possibility is using a Learning Analytics Cockpit in which students can directly monitor their activity during the whole course. Currently we are developing a prototype of a LA Cockpit.

References

[1] Three Megatrends That Will Drive Digital Business Into the Next Decade Cycle, Gartner, [Online]. Available: http://www.gartner.com/newsroom/id/3784363.

[2] T. Mitchell, E. Brynjolfsson, [Online]. Available: http://www.cs.cmu.edu/~tom/pubs/Nature2017_Mitchell_Brynjolfsson_FINAL.pdf , [20.04.2017].

[3] A. McAfee, E.Brynjolfson, "Big Data. The Management Revolution", Harvard Bussiness Review, 61-67, 2012. [Online]. Available: http://www.tias.edu/docs/default-source/Kennisartikelen/mcafeebrynjolfson_bigdatamanagementrevolution_hbr2012.pdf .

[4] M. Ciolacu, P. Svasta, W. Berg, H. Popp, "Education 4.0 for Tall Thin Engineer in Data Driven Society", IEEE- 23nd International Symposium SIITME, Constanta, Romania, 2017, in press.

[5] "What's the difference between learning techniques", Focus on Automotive, Electronic Design Library, [Online]. Available: http://www.electronicdesign.com/automotive/what-s-difference-between-machine-learning-techniques, [27.04.2017].

[6] A. Ng. , [Online]. Available: http://www.andrewng.org/publications/

[7] M. Ciolacu, R.Beer, "Adaptive user interface for higher education based on web technology", IEEE- 22nd International Symposium SIITME, DOI: 10.1109/SIITME.2016.7777299 , 2016.

[8] R. Ferguson, A. Brasher, "Learning Analytics: Vision of the Future", LAK ´16. Edinburgh, ACM.

[9] A. Samuel. "Some Studies in Machine learning Using the Game of Checkers", IBM Journal of Research and Development. 3, 1959, DOI:10.1147/rd.33.0210.

[10] Martin J. Cochran, "Cryptographic hash functions", Doctoral Dissertation, University of Colorado at Boulder Boulder, CO, USA, 2008, ISBN: 978-0-549-50843-4.

[11] A. F. Tehrani, "Learning Nonlinear Monotone Classifiers Using The Choquet Integral", Philipps University Marburg , dissertation, 2014.

[12] T. Mitchell, [Online]. Available: http://www.cs.cmu.edu/%7Etom/pubs/MachineLearning.pdf, 1998.

[13] B. Schölkopf, A. J. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond". MIT Press, Cambridge, MA, USA, 2001.

[14] J. R. Quinlan, "Induction of decision trees", Mach. Learn., 1(1):81–106, March 1986.

[15] P. Vasudevan. "Iterative dichotomiser-3 algorithm in data mining applied to diabetes database", Journal of Computer Science, 10(7):1151–1155, 2014.

[16] P. J. Rousseeuw, L. Kaufman. „Finding Groups in Data: An Introduction to Cluster Analysis", 2005, New Jersey.

[17] Keras Documentation, "The Python Deep Learning Library", [Online]. Available: https://keras.io .

[18] D. P. Kingma, J. Ba. "Adam: A Method for Stochastic Optimization", [Online]. Available: https://arxiv.org/abs/1412.6980 .

[19] A. Yadav, S. Dhingra. "A review on K-Means Clustering Technique", International Journal of Latest Research in Science and Technology, 2016, p.13-16.

[20] H. Popp, R. Beer. „Evaluation virtueller Mathematik-Kurse – Learnszenarienvergleich und Learning Analytics", in: „Evaluierung offener Lernszenarien", FH Joanneum, 2014.