International Journal of Applied Sciences and Smart Technologies
Volume 1, Issue 1, pages 65–82
ISSN 2655-8564

Factors Influencing the Difficulty Level of the Subject: Machine Learning Technique Approaches

Hari Suparwito

Department of Informatics, Faculty of Science and Technology,
Sanata Dharma University, Yogyakarta, Indonesia
Corresponding Author: shirsj@jesuits.net

(Received 07-05-2019; Revised 21-05-2019; Accepted 21-05-2019)

Abstract

The difficulty level of a subject is needed either to understand the student acceptance of the subject and the highest level of student achievement in it. Some factors are considered, what kind of instructions, the readiness of the instructor and students in teaching and learning, evaluation and monitoring systems, and student expectations. Many factors are involved, and educators should know this. It is better if they can discern which are the prime factors and which the secondary factors. The purpose of the study is to find out the determinant factors in establishing the difficulty level of the subject from the students', teachers' and infrastructure point of view using three machine learning techniques. The MSE and the variable importance measurement were used to predict between some factors such as Attendance, Instructors, and other factors as independent variables and the difficulty level of the subject as a dependent variable. The study result showed that Gradient Boosting Machine obtained the MSE value result 1.14 and 1.30 for training and validation dataset. The model generated five variable importance as an independent factor, i.e. Attendance, Instructor, The course can give a new perspective to students, The quizzes, assignments, projects and exams

Volume 1, Issue 1, pages 65–82
ISSN 2655-8564

contributed to helping the learning, and The Instructor was committed to the course and was understandable. The Gradient Boosting Machine is superior to other methods with the lowest MSE and MAE values results. Two methods, Gradient Boosting Machine and Deep Learning, have produced the same five main factors that influenced the difficulty of the subject. It means these factors are significant and should get intention by the stakeholders

Keywords: machine learning, regression, deep learning, random forest, gradient boosting machine, data mining, education.

1 Introduction

Education provides people with knowledge about life and the world. It helps build character and leads to illumination. Given the importance of education, researchers ask themselves what factors influence the process of teaching and the attitude of students so that the students can understand the subjects, and what factors help to measure the difficulty level of subjects. The difficulty level of subjects is needed both to understand either the student acceptance of their subject or to ascertain the highest level of the student achievement in them [1]

John D. et al. [2] have examined some aspects and conducted some reviews based on learning conditions, student characteristics, materials and criterion tasks for effective learning techniques. Another group of researchers [3] have found that the social context influenced effective teaching and learning. Some factors mentioned were direct instruction, frequent monitoring, sense of communities, and student expectations. There are many factors involve here.

Research on education using data mining are increasing and promising in the last years and mostly focusing the research on student's performance, the effectiveness of learning and students and teacher's perception of learning [4]. Romero et al. stated that the objective using data mining in education areas is to improve the learning itself and the actors are students and teachers with the subjects of learning and the way to deliver as a medium relates them. Vanthienen and De Witte [5] revealed that their study

Volume 1, Issue 1, pages 65–82
ISSN 2655-8564

showed the use of machine learning methods is advantageous especially when it faces a nonlinear interaction function such as the role of a school principal to accommodate the district size policies. Another research in education field using the machine learning technique was undertaken by Liao, Zingaro [6]. They stated that using machine learning techniques; they can identify students who are at risk of performing poorly in a course.

Moreover, the machine learning approach was also performed for evaluating and predicting the student's level of proficiency [7]. To successfully predict the quality of this type of educational process the authors use one of the machine learning techniques. They claimed that the proposed technique could be effectively used in the educational management when the online teaching strategy should be selected based on student's goals, individual features, needs and preferences. Finally, Cope and Kalantzis [8] claimed that the use of machine learning and big data analysis in research on education should be undertaken because these emerging sources of evidence of learning have significant implications for the relationships between assessment and instruction. Moreover, for educational researchers, these datasets are in some senses different from conventional evidentiary sources, and this raises a new approach and give a different point of view to the traditional research in education areas.

The objective of this research is to find out the determinant factors that affect the student's acceptance focusing on the difficulty level of students understanding of the subjects. Instead of using a statistical approach in this present study we performed three machine learning techniques, i.e. Deep Learning, Random Forest, and Gradient Boosting Machine. Another purpose of this research is to introduce and compare the results of three machine learning methods in education areas. As the data set, we collected the dataset from the student evaluation at Gazi University Ankara [9] and was taken from the UCI repository dataset. This data set will be examined by three machine learning techniques using H2O platforms.

This paper is organised as follows. In section 2, we describe the research methodology with the following process in data mining approaches and then the results based on the H2O data mining tools calculation are presented and discussed in section 3. In chapter 4, we provide the conclusion and the subsequent work research outcome.

International Journal of Applied Sciences and Smart Technologies

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

2 Research Methodology

In general, the steps in this study follows the model of data mining techniques [10]:

2.1 Objective determination

The first step was to discover the real-world problems. This study will attempt to answer the educational question of how to understand and measure the difficulty level of the subject from the students', teachers' and infrastructures' point of view. To be more precise, the following research question was raised: What is the determinant factors which make students think and establish that this subject is difficult or easy?

A hypothesis was created to test which attributes in the data set gives a significant contribution toward the research question: Students think that the level of the subject difficulty is more likely to be influenced by the subject syllabus, activities and interactions between students and instructors and the readiness of students and teachers to engage in the learning process.

By analysing and testing this hypothesis, it shall know the determinant factors to answer the question of why do the students think that the subject is difficult to understand? Moreover, what should be done by the teachers so that the students can accept and understand the subject materials more easily?

2.2 The proposed work

To examine three machine learning models we selected the dataset from the UCI machine learning repository about Turkiye Students Evaluation data set [9]. Furthermore, the dataset was analysed for reducing its dimensional features by using Principle Component Analysis (PCA) and then followed by performing a data normalisation using z-normalization. Moreover, the dataset was randomly split into ratio 80% : 20% from the data population as training and validation dataset.

Three machine learning techniques then were applied to training dataset obtaining the regression model, the MSE and MAE value results and the variable importance of each method. Using the model, we observed validation dataset to find out the MSE and

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

MAE values results and the variable significance for the testing dataset. All processes can be seen in the diagram below.

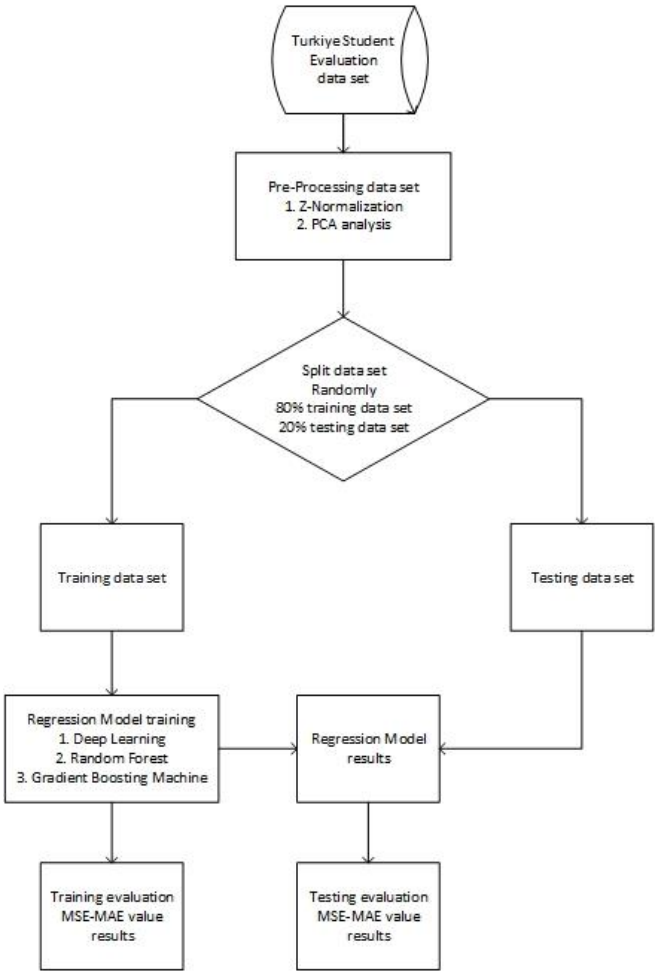

Figure 1. The proposed work. It was started by selecting the dataset to deliver the MSE and MAE value results and the variable importance rank

2.3 Data pre-processing

The research used the data result of the student questionnaire at Gazi University Ankara Turkey [9]. The dataset was obtained from the UCI machine learning repository dataset (https://archive.ics.uci.edu/ml/index.php). There are 5820 instances in the data set with 33 attributes where 28 attributes are formed in a Likert-type scale with the value from 1 to 5. The Likert-type scale values 1 equals to a strongly disagree value, and the value 5 equals to a strongly agree value. The five other attributes are questions with the answers in the natural numbers data format. The questions can be grouped into

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

three substantial group questions based on students', teachers' and infrastructures' point of view.

Next, we undertook a PCA analysis for features reduction. Matrix correlation from the PCA analysis showed each eigenvalue of the features. A new variable (principal component) was calculated based on eigenvalues with the values bigger than one. The PCA analysis result for a new variable is five principal components. We analysed and found that five principles components can be grouped into Attendance, Instructor, subject preparation, quizzes or exams, and the relationship between students and instructors.

Table 1. Table of principle component

Component	Standard Deviation	Proportion of Variance	Cumulative of Variance
PC1	6.140	0.588	0.588
PC2	3.686	0.212	0.800
PC3	1.701	0.045	0.845
PC4	1.411	0.031	0.876
PC5	1.059	0.017	0.894

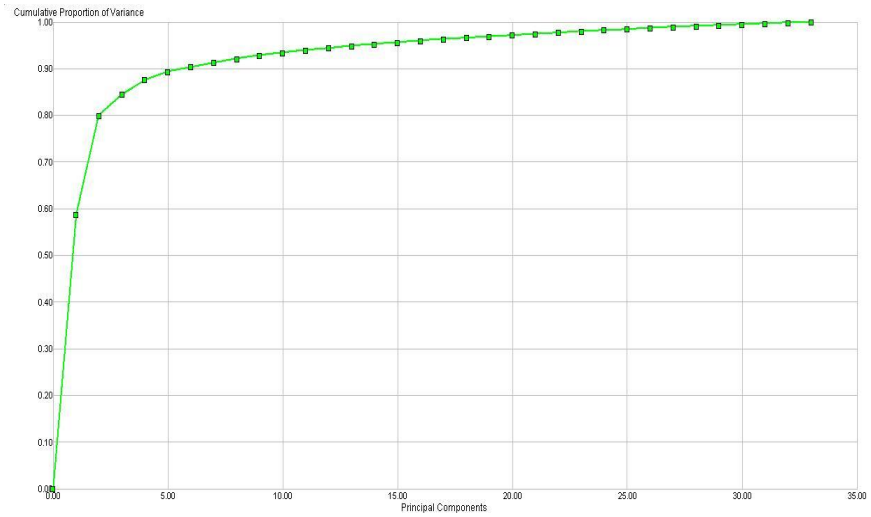

Figure 2. The cumulative proportion of variance versus principle component.

From five principal components, we selected which features have a high rank based on the eigenvector values of each feature. Finally, we found 15 features that can be used

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

in this study. Therefore, the number of features was reduced from 33 features to 15 features only. A new reduced feature is shown in the following table.

Table 2. PCA analysis results

Features	Name of features
	Difficulty (target label)
	Attendance
	Instructors
Q1	The semester course content, teaching method and evaluation system were provided at the start
Q4	The course was taught according to the syllabus announced on the first day of class.
Q5	The class discussions, homework assignments, applications and studies were satisfactory.
Q7	The course allowed fieldwork, applications, laboratory, discussion and other studies.
Q8	The quizzes, assignments, projects and exams contributed to help the learning.
Q12	The course helped me look at life and the world with a new perspective.
Q16	The Instructor was committed to the course and was understandable.
Q21	The Instructor demonstrated a positive approach to students.
Q22	The Instructor was open and respectful of the views of students about the course
Q24	The Instructor gave relevant homework assignments/projects and helped/guided students.
Q25	The Instructor responded to questions about the course inside and outside of the course.
Q27	The Instructor provided solutions to exams and discussed them with students.
Q28	The Instructor treated all students in a right and objective manner.

2.4 Data mining

The next process after the data pre-processing was to decide the kind of evaluation to be applied to the data set. The regression task is chosen because the data set is already classified in attributes and the questionnaire's answer is on a Likert-type scale from 1 to 5 means already classified too. Another reason is that this study' goal is directed to discover which attributes are the determinant factors of the difficulty level of the subject.

Three machine learning techniques that are Deep Learning (DL), Random Forest (RF) and Gradient Boosting Machine (GBM) were used to examine the data set focusing on the regression analysis between 15 attributes as an independent variable and the difficulty level of the subject as a target or dependent variable.

2.4.1. Deep Learning

Introduced the first time by Hinton et al. DL becomes more and more popular as one method to solve the problems in machine learning areas [11]. Deep learning is a part of machine learning techniques that aim to imitate the work of the human brain using an

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

artificial neural network. Different from other machine learning programs, the deep learning algorithm is made by a complex and high capability to learn, work and classify data.

In general, DL consist of 3 main layers: input-hidden-output. Input layers work for containing raw data as input data. Hidden layers are applied for observing, learning and classifying data based on the references, in case of DL hidden layers usually consist of more than three layers. Output layers present the results.

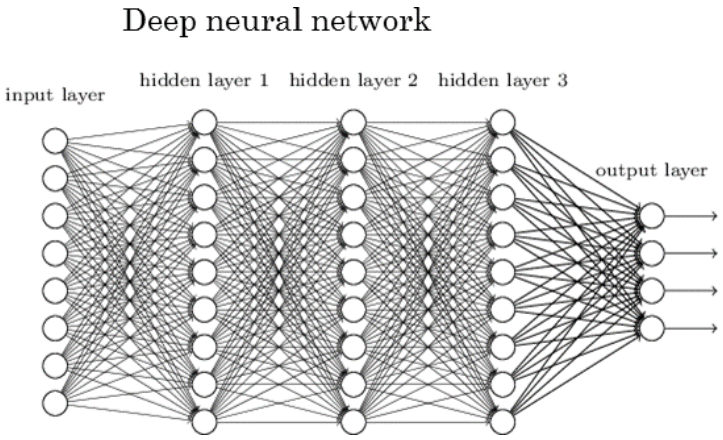

Figure 3. Deep Learning diagram (the picture was taken from https://www.kdnuggets.com/2017/05/deep-learning-big-deal.html).

2.4.2. Random Forest

Random Forest is an ensemble learning technique for classification [12]. RF works by constructing a collection of decision tree at training time and returning the class that is the mode of all of the classes of the individual trees. Like DL, the RF algorithm has a significant advantage when analysing many of the datasets. It can address high-dimensional data with an excellent ability to learn from a large amount of data, and it can realise learning regression and classification for nonlinear sample data.

International Journal of Applied Sciences and Smart Technologies

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

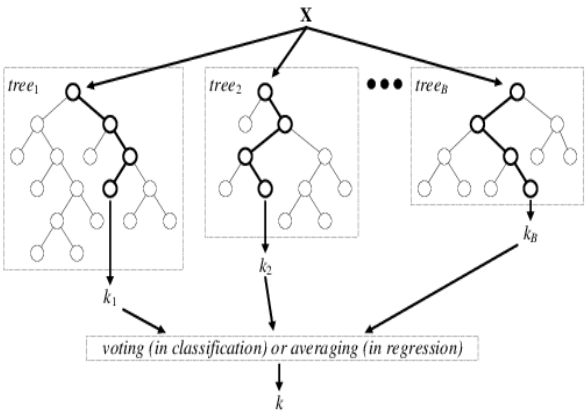

Figure 4. Random Forest architecture for classification and regression analysis (picture was taken from https://www.researchgate.net/figure/Architecture-of-the-random-forest-model_fig1_301638643)

2.4.3. Gradient Boosting Machine

Gradient boosting is a form of machine learning boosting. Boosting means target outcomes for each case are set based on the gradient of the error to the prediction. The idea behind GBM is to set the target outcomes for the next model in order to minimise the error. Each new model performs in the direction that minimises prediction error [13]. Even though RF and GBM are an ensemble learning method, GBM and RF differ in the way the trees are created: the order and the way the results are combined. GBM tries to add new trees that compliment the already built ones. This usually gives a better accuracy with fewer trees. Therefore, GBM performs better than RF if parameters tuned carefully [14].

2.4.4 Cross-Validation

The goal of cross-validation is to test the model's ability to predict new data and to give an insight into how the model will generalise to an independent dataset. In each machine learning model was undertaken the K-fold Cross-Validation (CV) method and it was applied to training and testing data set. The K-fold CV method was selected for the data sampling method because data instances should be evaluated in training and testing data set. The number of instances is quite large so when the K-fold CV does the data sampling to the training and testing data set K-fold CV can do quite well. This

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

experiment was repeated many times, in this case, the repeating times was expressed by the K values. Even for some scientists argued that K=10 is the best value but in this research, the selection of the best K value in K-fold CV done by repeating many times experiment using various K values [15]. In this study, K-fold CV equal to 10 was applied.

Machine learning methods worked by using some parameters and finding the best result, each machine learning method has specific parameters to adjust. We used data grid analysis to find the best parameters to provide the optimum results. The following table shows the grid search parameters applied for

Table 3. Grid parameters values model

Model	Grid Parameter values
DL	Function – Rectifier; Tanh
	Hidden layers – 200, 200, 100, 50; 100,100,50; 50, 100, 100, 50
	Epochs – 50; 100; 200
	CV – 5; 10
RF	nTrees – 50; 100; 200
	Epochs – 50; 100; 200
	CV – 5; 10
GBM	nTrees – 50; 100; 200
	Epochs – 50; 100; 200
	CV – 5; 10

The best performance from each model showed by the following parameters

Table 4. Parameters values model

Model	Parameter values
DL	Function – Rectifier
	Hidden layers – 200, 200, 100, 50
	Epochs – 200
	CV – 10
	Input dropout – 0.2
RF	nTrees – 200
	Epochs – 100
	CV – 10
GBM	nTrees – 50
	Epochs – 50
	CV – 10

Tabel 4 shows the best parameters gave by the grid search analysis.

International Journal of Applied Sciences and Smart Technologies

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

3 Results and Discussions

Three machine learning methods were used to examine the dataset. The results obtained were the MSE and MAE values of each method and the variable importance. The Mean Squared Error (MSE) value was used to find the difference between the estimator and what is estimated. The MSE is achieved by applying the following formula:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1}$$

Where $\check{Y}$ is a vector of n prediction and Y is the vector of observed values corresponding to the input to the function that created the predictions. Y_i is the i-th value of the vector.

In this study, the training dataset was the data obtained from 80% number of data population, while the dataset from the rest of the number of populations (20%) was used as a testing dataset. H2O machine learning tools were performed for training and testing dataset, and the MSE value results are presented in the following table.

Table 5. MSE and MAE values of three machine learning models

Models	Training data set		Validation data set	
	MSE	MAE	MSE	MAE
DL	1.25	0.89	1.33	0.92
RF	1.31	0.92	1.38	0.91
GBM	1.14	0.84	1.30	0.90

The lowest MSE values are the best result because it describes the similarity between the real values and the prediction values. In other words, the lower the MSE, the higher the accuracy of prediction as there would be an excellent match between the actual and predicted data set. In this study, the lowest MSE value is obtained by GBM models.
Like the MSE value, the MAE value obtained by the formula

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n} \tag{2}$$

Where x and y values are observed and predicted values. The lower MAE value also indicates better performance of the models.

International Journal of Applied Sciences and Smart Technologies

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

Understanding the best model for the prediction can be performed by using deviance of training and testing dataset [16]. Deviance measurement is used for measure how well the model to predict It attempt is a generalisation of the idea of using the sum of squares of residuals in ordinary least square to cases where model-fitting is obtained by maximum likelihood. The following picture shows the deviance score for each number of trees in GBM.

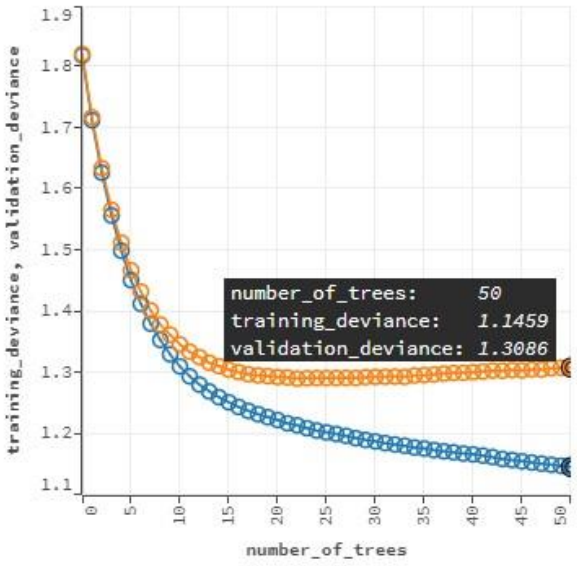

Figure 5. GBM deviance score for each number of trees. We show the GBM model result only because GBM method obtained the best result

3.1. Variable importance

Wei, Lu [17] stated that it is essential to know which the more significant factor or variable in the regression or prediction analysis. Whereas Grömping [18] argued that predictive analysis would be more convincing when the most influential predictor variable obtained, though the way to find variable importance is challenging and some regression models are not directly planned to find the variable importance. Therefore, another method needs to be used to find the variable importance. Some techniques in machine learning could be used as an alternative way to find the variable importance, especially when dealing with high-dimensional input data and the categorical output.

Which variables are more significant in predicting the difficulty of the subject? Three ML methods were applied in this study. The percentage of Mean Square Error (MSE)

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

and Mean Absolute Error (MAE) was measured, which indicates which variable has a more significant influence compared with other variables in predicting the difficulty of the subject values. Table 6 shows the rank of the variable importance results and it also is given for example the graph of the variable importance from the GBM result in fig. 6

Table 6. variable importance results of each models

Models	Variable importance
DL	1. Attendance
	2. Instructure
	3. Q12 - The course helped me look at life and the world with a new perspective.
	4. Q16 - The Instructor was committed to the course and was understandable
	5. Q8 - The quizzes, assignments, projects and exams contributed to help the learning.
RF	1. Attendance
	2. Q22 - The Instructor was open and respectful of the views of students about the course.
	3. Q25 - The Instructor responded to questions about the course inside and outside of the course.
	4. Q21 - The Instructor demonstrated a positive approach to students.
	5. Instructure
GBM	1. Attendance
	2. Instructure
	3. Q12 - The course helped me look at life and the world with a new perspective.
	4. Q8 - The quizzes, assignments, projects and exams contributed to help the learning.
	5. Q16 - The Instructor was committed to the course and was understandable.

DL and GBM models have the same variable importance even though for Q8, Q12 and Q16 have a different rank. However, the main five factors are the same that was produced by DL and GBM analysis. For three machine learning models, two main factors are attendance and instructors have a significant influence in determining the difficulty level of the subject. It means these two factors are the most important predictor for the difficulty of the subject variable.

The previous study also revealed that student's performance was not only dependent on their academic effort but also some other aspect that has a similar influence as well [19].

International Journal of Applied Sciences and Smart Technologies

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

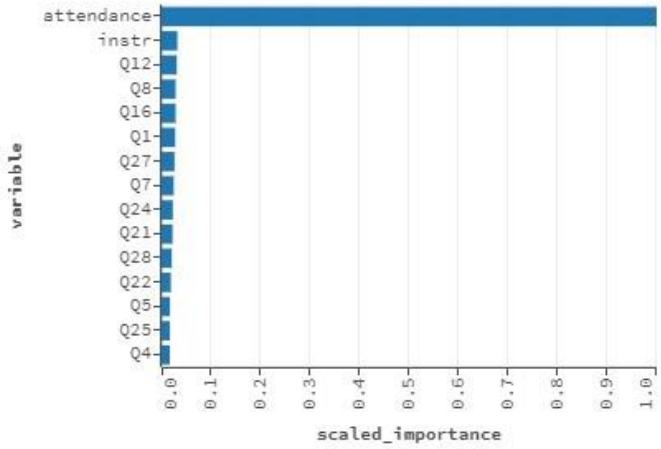

Figure 6. GBM variable importance

To answer the main question in the first section, now we can see the rank of the variable importance, especially from DL and GBM results. Moreover, if we observe which features have a significant influence, we can draw some points here,

a) Attendance has the most significant impact. The respondent thought that attendance whether by students or by instructors have an important role and it can make their presumption about the subjects. Attendance means participation and involvement between students and instructors.

b) Instructors and their attitudes or approach to the students are related to the subjects. The students are convinced that the instructors have a significant impact on delivering the subjects to them whether it was easy or difficult to be understood by them. This aspect is also related to the instructors' attitude such as how the instructor was committed to the course, how they respond if students are asking the subject in or out classes, how they can encourage the students to do the best with the selected subjects. The previous study by Martin, Wang [20] stated that instructors become an essential factor to make the subjects were easy or difficult in front of their students.

c) The course can give a new perspective to students. A new perspective could be driven by the students. Therefore, they would focus on learning the subject and the next it will make the subject was easy to learn. In other words, giving a new

Volume 1, Issue 1, pages 65–82
ISSN 2655-8564

perspective for life become a stimulus to the students to learn and love the subjects.

d) The quizzes, assignments, projects and exams contributed to help the learning. The students need the way to express their ability in understanding the subjects. The students felt that reading some theories were not enough, they needed some exercises, and by doing the exercises, they can understand the subject more. These aspects were also mentioned by Henderson and Harper [21] in their research. They revealed that some correction, assessment, and teacher's feedback on student's quizzes could help the students to prepare their exams better.

4 Conclusions

Three machine learning algorithms, i.e. Deep Learning, Random Forest, and Gradient Boosting Machine with K-folds CV data sampling methods have been applied to analyse the difficulty level of the subject based on students', teachers' and infrastructures' point of view. The data set is collected from the student questionnaire result at Gazi University Ankara. The result revealed that there are five determinant factors, i.e. Attendance, Instructors, the course helped me look at life and the world with a new perspective, the quizzes, assignments, projects and exams contributed to helping the learning, and the Instructor was committed to the course and was understandable. These five determinant factors can affect student's and instructor's perspective on the difficulty level of the subject. The two main factors are Attendance and Instructors. This study also demonstrated that data mining methods could be employed in the education field. However, the ability to understand data and how to work with them is very crucial. Data mining processes are important especially step by step at the stage model of data mining can be used as guidance on how to work with the data mining to solve the real-world problems.

In the subsequent study, it is possible to discover and compare these techniques with another algorithm in classification and regression tasks. Another possibility is also to compare some other tools such as Orange and Rapidminer tools where these two tools work on machine learning algorithm for solving the same problem.

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

Acknowledgements

This research was supported by Department of Informatics Engineering, Sanata Dharma University. We would also like to thank the anonymous reviewers; whose comments greatly improved the manuscript.

References

[1] M. T. Tillery and A. Fishbach, "How to measure motivation: a guide for experimental social psychologist," *Social and Personality Psychology Compass*, **8** (7), 328−341, 2014.

[2] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham, "Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology," *Psychological Science in the Public Interest,* **14** (1), 4−5, 2013.

[3] P. Hallinger and J. F. Murphy, "The social context of effective schools," *American Journal of Education,* **94** (3), 328–355.

[4] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* **3** (1), 12−27, 2013.

[5] J. Vanthienen and K. D. Witte, "Data analytics applications in education," *CRC Press Taylor & Francis Group,* 2017.

[6] Liao, S.N., et al., "A robust machine learning technique to predict low-performing students," *ACM Transactions on Computing Education (TOCE),* **19** (3), 18, 2019.

[7] N. Kushik, N. Yevtushenko, and T. Evtushenko, "Novel machine learning technique for predicting teaching strategy effectiveness," *International Journal of Information Management,* (2016). https://doi.org/10.1016/j.ijinfomgt.2016.02.006

[8] B. Cope and M. Kalantzis, "Big data comes to school: implications for learning, assessment, and research," *AERA Open,* **2** (2), 1–19, 2016.

[9] G. Gunduza and E. Fokoue, *Turkiye student evaluation I,* University of California, School of Information and Computer Sciences, 2013.

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

[10] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi, "Discovering data mining: from concept to implementation," *Englewood Cliffs, N. J. Prentice Hall,* 1998.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* **521** (7553), 436−444, 2015.

[12] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news,* **2** (3),18−22, 2002.

[13] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics,* **29** (5), 1189−1232, 2001.

[14] R. E. Schapire, "The boosting approach to machine learning: an overview, in nonlinear estimation and classification," *Springer,* 149−171, 2003.

[15] R, Kohavi and F. Provost, "Confusion matrix," *Machine Learning,* **30** (2-3), 271−274, 1998.

[16] G. Ritschard, "Computing and using the deviance with classification trees," *COMPSTAT 2006-Proceedings in Computational Statistics,* 55−66, August 2006.

[17] P. Wei, Z. Lu, and J. Song, "Variable importance analysis: a comprehensive review," *Reliability Engineering & System Safety,* **142,** 399−432, 2015.

[18] U. Grömping, "Variable importance in regression models," *Wiley Interdisciplinary Reviews: Computational Statistics,* **7** (2), 137−152, 2015.

[19] A. A. Saa, "Educational data mining & students' performance prediction," *International Journal of Advanced Computer Science and Applications,* **7** (5), 212−220, 2016.

[20] F. Martin, C. Wang, and A. Sadaf, "Student perception of helpfulness of facilitation strategies that enhance instructor presence, connectedness, engagement and learning in online courses," *The Internet and Higher Education,* **37**, 52−65, 2018.

[21] C. Henderson and K. A. Harper, "Quiz corrections: improving learning by encouraging students to reflect on their mistakes," *The Physics Teacher,* **47** (9), 581−586, 2009.

International Journal of Applied Sciences and Smart Technologies

Volume 1, Issue 1, pages 65–82

ISSN 2655-8564

This page intentionally left blank