Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi/Gazi University Journal of Gazi Educational Faculty

GEFAD

ISSN: 1301-9058 GEFAD / GUJGEF 45(2): 543-568(2025)

Investigation of Prediction Accuracy of Hierarchical Linear Modelling and Artificial Neural Networks Methods on PISA 2018 Reading Literacy

Hiyerarşik Doğrusal Modellemenin ve Yapay Sinir Ağları Yöntemlerinin PISA 2018 Okuma Okuryazarlığı Tahmin Doğruluğunun Araştırılması

Eda Akdoğdu Yıldız[1*], Kübra Atalay Kabasakal[2]

[1]Australian Council for Educational Research, Australia

e-mail: eda.yildiz@acer.org, ORCID ID: https://orcid.org/0000-0003-4374-4379

[2]Hacettepe University, Turkiye

e-mail: katalay@hacettepe.edu.tr, ORCID ID :https://orcid.org/0000-0002-3580-5568

Article Information

Article Type: Research Article

Received: 17.05.2025

Accepted: 08.08.2025

Year: 2025 | Volume: 45 | Issue: 2 | Pages: 543-568

[*]**Corresponding Author:** Eda Akdoğdu Yıldız

Author Note: This study was produced from the first author's dissertation thesis.

Cite as: Akdoğdu Yıldız, E., & Atalay Kabasakal, K. (2025). Investigation of prediction accuracy of hierarchical linear modelling and artificial neural networks methods on PISA 2018 reading literacy. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi [Gazi University Journal of Gazi Educational Faculty]*, *45*(2), 543-568.

Ethics Committee Approval: Not applicable. This study did not involve human participants, identifiable personal data, or any interventions, and was based entirely on secondary analysis of existing materials; therefore, review and approval by an institutional ethics committee was not required.

Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi
Gazi University Journal of Gazi Educational Faculty
https://dergipark.org.tr/tr/pub/gefad

Investigation of Prediction Accuracy of Hierarchical Linear Modelling and Artificial Neural Networks Methods on PISA 2018 Reading Literacy

ABSTRACT

In this research, it is aimed to compare hierarchical linear modelling and artificial neural network estimation methods in predicting students' reading comprehension success in the Program for International Student Assessment (PISA) 2018 application. In accordance with this purpose, it is planned to determine how students' PISA success status is estimated at student and school level, the data mining method used in estimation and the explained variance and error values of multilevel modelling. The type of study is, in a way, relational research because of the establishment of models in which there are relationships between dependent and independent variables. On the other hand, it is descriptive research in terms of performing analyses with two methods for each country sampled in the study and comparing the results obtained in terms of explained variance and error values. In this research, the performance of data mining techniques (artificial neural networks – ANN) and multilevel analysis methods (hierarchical linear modeling – HLM) in the field of education is evaluated. It has been determined that HLM carries out the estimation process with lower error and higher R^2 than ANN in the analysis of multi-level data. In addition, HLM provides more information about the predictive level of the variables and the variance that is not explained by the variables in the model compared to ANN. For this reason, HLM analysis was used to examine the variables that affect reading comprehension success in the study. As a result, it was seen that the student level and school level variables added to the model had a statistically significant effect on reading comprehension achievement. While teacher-directed instruction and lack of educational material at school cause negative effects on reading comprehension success, it has been determined that economic-social-cultural situation, metacognitive strategies, disciplinary climate in the classroom, teacher support, and staff shortage variables have positive effects. The results obtained are generally in agreement with similar studies in the literature.

Keywords: *Hierarchical linear modelling, Data mining, Artificial neural networks, Reading comprehension, PISA*

ÖZ

Bu araştırmada Uluslararası Öğrenci Değerlendirme Programı (PISA) 2018 uygulamasında öğrencilerin okuduğunu anlama başarısını tahmin etmede hiyerarşik lineer modelleme ve yapay sinir ağları tahmin yöntemlerinin karşılaştırılması amaçlanmaktadır. Bu amaç doğrultusunda; öğrencilerin PISA başarı durumlarının birey ve okul düzeyinde nasıl tahmin edildiği, tahmin etmede kullanılan veri madenciliği yöntemi ve çok düzeyli modellemenin açıklanan varyans ve hata değerlerinin belirlenmesi planlanmaktadır. Çalışmanın türü, bağımlı ve bağımsız değişkenler arasında ilişkilerin bulunduğu modellerin kurulmasından dolayı bir yönüyle ilişkisel araştırmadır. Diğer bir yönüyle ise çalışmada örnekleme alınan her ülke için iki yöntemle analizler gerçekleştirilip elde edilen sonuçların açıklanan varyans ve hata değerleri açısından karşılaştırılması bakımından betimsel araştırma niteliğindedir. Bu araştırmada eğitim alanında da kullanılmaya başlanan veri madenciliği (yapay sinir ağları-YSA) ve çok düzeyli analiz yöntemlerinin (hiyerarşik lineer model-HLM) nasıl performans gösterdiğine ilişkin bulgular elde edilmiştir. HLM'in çok düzeyleri verilerin analizinde YSA'ya göre daha düşük hata ve daha yüksek R^2 ile tahminleme sürecini yürüttüğü belirlenmiştir. Ayrıca HLM değişkenlerin yordama düzeyi ve modelde yer alan değişkenler tarafından açıklanmayan varyans hakkında YSA'ya göre daha fazla bilgi sunmaktadır. Bu sebeple çalışmada okuduğunu anlama başarısını etkileyen değişkenleri incelemek için HLM analizi kullanılmıştır. Sonuç olarak

modele eklenen birey düzeyi (düzey 1) ve okul düzeyi (düzey 2) değişkenlerinin okuduğunu anlama başarısı üzerinde istatiksel olarak anlamlı etkisi olduğu görülmüştür. Okuduğunu anlama başarısı üzerinde öğretmen yönlendirmeli öğretim ve okuldaki eğitimsel materyal eksikliği negatif yönlü etkiye sebep olurken ekonomik-sosyal-kültürel durum, üstbiliş stratejileri, sınıftaki disiplin iklimi, öğretmen desteği, personel eksikliği değişkenleri pozitif yönlü etkisi olduğu tespit edilmiştir. Elde edilen sonuçlar genel olarak literatürde yer alan benzer çalışmalarla uyum göstermektedir.

Anahtar Sözcükler: *Hiyerarşik lineer modelleme, Veri madenciliği, Yapay sinir ağları, Okuduğunu anlama, PISA*

INTRODUCTION

The role of acquiring, processing, and evaluating information is increasingly emphasized due to its potential contributions to technological, scientific, and economic advancements (Akkoyunlu, 2008). Among the essential language skills, reading holds a crucial position as a primary means of gaining knowledge. Reading is the process of deriving meaning from written texts, a complex skill that involves coordinating interrelated information sources (Anderson et al., 1988). The outcome of this process is learning, underscoring the importance of comprehension. Merely decoding numbers and symbols is insufficient for effective information processing and evaluation (Çelenk, 2003). Reading literacy involves extracting meaning and understanding details within the text, allowing readers to logically structure the message provided (Radoyevic, 2006; Rose et al., 2000). This skill can be developed at any life stage, beginning in elementary school. Identifying students' reading literacy levels and developing targeted programs for those needing improvement is therefore essential.

Reading literacy is a crucial factor for students to achieve their desired goals in daily and academic life. Various studies have emphasized that reading literacy involves not only comprehending text but also the ability to analyze, reflect, and make connections within texts to support lifelong learning (Guthrie & Wigfield, 2000; Organisation for Economic Co-operation and Development [OECD], 2019; Snow, 2002). According to PISA, reading literacy encompasses a student's ability to understand, utilize, evaluate, associate, and reflect on texts presented in various formats to attain their objectives, expand their knowledge, develop their potential, and engage in society (OECD, 2019).

When employing PISA data to investigate the factors influencing reading literacy, the students are assessed in terms of affective and cognitive measures, and they are nested within hierarchical organizational structures such as classes, schools, regions, and countries. Data from large-scale international assessments like PISA contain substantial depth and breadth, owing to their multilevel structure and extensive scale. In analyzing such hierarchical data structures, the use of data science, which has seen rapid growth in its application to the field of education, can be highly beneficial, as it is in various other domains such as economics, informatics, and marketing.

Data science encompasses a range of advanced analytical approaches, including data mining, machine learning, deep learning, and statistical modelling. These techniques are utilized to derive meaningful insights from data sets and translate them into actionable business strategies (Ali et al., 2023). One such analytical method is artificial neural networks (ANNs), which are widely employed today. ANNs have demonstrated continuous progress and improvement in their extraction rules and training algorithms, enabling them to perform effectively even with inconsistent or erroneous data, and achieve low error rates in their outputs (Ni, 2008).

Researchers require the information provided by the analyses they choose to employ in their study, as well as the similarities and differences of this information compared to other analyses. Consequently, it is essential to compare the analyses for analogous purposes. This study compared approaches that are suitable for usage in hierarchical data, such as artificial neural networks, with hierarchical linear modelling in order to determine the predictive power of variables affecting students' reading literacy. The study aimed to determine how students' PISA scores are predicted at both the student and school levels. It also sought to identify the similarities and disparities in the evaluation metrics, demonstrating the predictive success of HLM and ANNs. Aligned with the study's objective, the findings hold multifaceted significance. This importance augments the study in two dimensions: identifying

the variables predicting reading literacy and comparing the suitable analyses for hierarchical data. One of the study's strengths was obtaining results from diverse perspectives and making recommendations accordingly.

HLM and ANNs analyses have advantages and disadvantages that differ from each other. HLM is particularly advantageous when data has a hierarchical structure with two or more levels, and it has been shown to handle unbalanced data and missingness more effectively compared to traditional methods, which may require additional techniques to address these issues (Raudenbush & Bryk, 2002). A nested structure is likely observed, especially in big data. Therefore, the use of HLM in such data is common. One of the most important differences distinguishing HLM from ANNs is that ANNs have no assumptions in analysing big data (Abiodun et al., 2018). The other advantages of ANNs are solving problems without mathematical modelling, solving nonlinear problems, learning from data and adapting what we have learned to new situations, and using it in real time. In addition to these, their disadvantages include shortage of certain rules for determining the network structure, the number of layers and cells, and the termination of the training process; shortage of good solutions when the training data does not contain good examples (Kumova & Kışla, 2020). Considering these differences, the aim of revealing whether the advantageous situations increase the predicting success in the analysis results increases the significance of the study.

Data mining/machine learning techniques have lately been utilised with regression approaches for prediction in the social sciences. In research examining the performance of prediction, ANN, one of the data mining algorithms, has been reported to perform better than known regression models (Lin et al., 2012). However, these results were mostly limited to single-level regression models. Few studies compared the analyses of multilevel data, such as HLM, with ANNs (Feng & Jones, 2015; Levy et al. 2020; Weng & Luo, 2023). Researchers have also been interested in whether ANNs can effectively model and predict outcomes in multilevel data, given their demonstrated performance on single-level datasets. Although data mining (DM) in education is a very new field of research, it is promising in identifying and meeting education needs. However, it is expected to present evidence emphasising the strengths of this field with many more studies for the improvement of the field. This study is increasingly important in meeting expectations, determining whether ANN, one of the DM methods, would be a strong alternative to HLM, which is frequently used in education.

This study aims to compare HLM, a methodology suitable for nested data structures, with ANN, a data mining technique, in order to assess the predictive power of variables influencing the reading comprehension performance of 15-year-old students in the PISA 2018. The comparative analysis of these approaches provides valuable insights into the performance of machine learning and other advanced statistical techniques recently applied in the field of education. Additionally, the study examines the variables that predict the reading literacy of 15-year-old students by focusing on three countries representing low, medium, and high reading literacy levels to increase the variability in the dependent and independent variables. This research will contribute to the understanding of the factors affecting reading literacy.

Hierarchical Linear Model

The data obtained from PISA have a hierarchical structure. Statistical and conceptual problems encountered in studies conducted with single-level modelling with hierarchical data have led to the development of multilevel modelling techniques. Multilevel models are useful and necessary to the extent that the data analysed provides sufficient variation at each level. Dividing variance into its intra-group and inter-group components provides evidence for researchers to decide whether a multilevel analysis is justified (Hox, 2010). Determining the degree of homogeneity between units in the same group will emphasise the dependence of observations on the group. The increase in the degree of homogeneity, the intraclass correlation, requires multilevel analysis. The necessity of multilevel analysis is decided by calculating the intraclass correlation, which is the ratio of the variance between groups to the total variance.

Researchers employing HLM commonly utilise a sequential model-building strategy, systematically testing and comparing increasingly complex nested models. This approach entails evaluating the four primary HLM stages in succession: the random effects ANOVA model, the means as outcomes model, the random intercepts model, and

the random intercepts and slopes model. The selection of the most suitable model is guided by statistical fit indices, theoretical justifications, and the specific research objectives at hand (Raudenbush & Bryk, 2002).

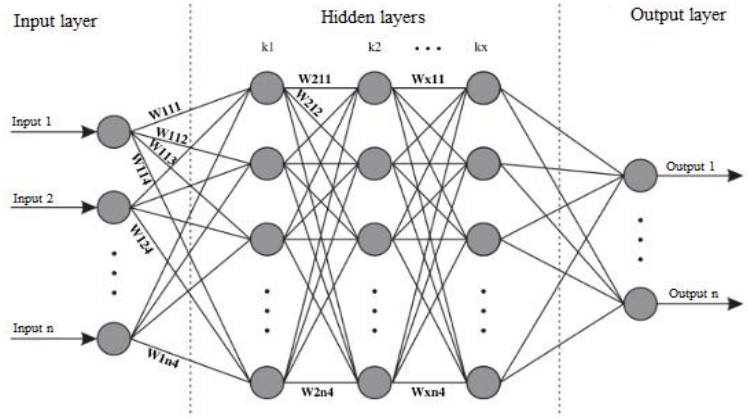
Artificial Neural Networks

Artificial neural networks are learning algorithms that simulate biological neural networks (Baba, 2024). Many predictions and classifications frequently use them due to their ability to "learn" from data, their non-parametric structure (shortage of strict assumptions), and their ability to generalise (Ni, 2008). In addition to generalisation, it offers solutions to a wide range of problems in various application areas due to its advantages of error tolerance, parallel operation, compatibility and non-assumption.

Neural networks use training datasets to recognise patterns of the overall data structure. They then predict the outputs for a new similar dataset. In this study, we partitioned the dataset, allocating 25% for training the model to identify patterns in the data, while reserving the remaining 75% for testing the model's ability to make accurate predictions on previously unseen information. As noted by Hastie et al. (2009), this approach helps to mitigate overfitting and provides a more reliable estimate of the model's real-world applicability.

Figure 1

Multilayer Artificial Neural Network Model

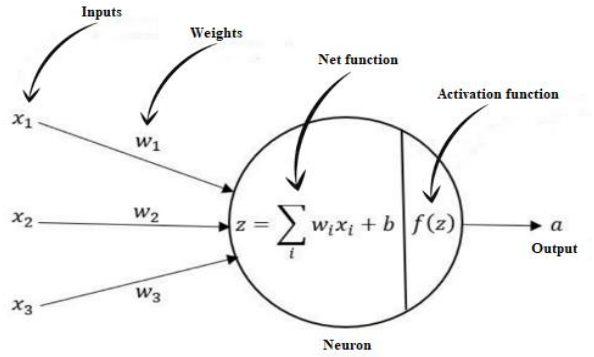


As in Figure 1, in the ANN, there are neurons and layers, which are the core processing units of the network. First is the input layer, where the algorithm includes the variables. This layer directs the information entering the system to the hidden layers without action. Hidden layers process the information and send it to the next hidden or output layer. In other words, there can be more than one hidden layer, and the information processed in the last hidden layer is redirected to the output layer. Since what happens in this layer cannot be fully explained, the hidden layer of the ANN is described as a black box. At the end is the output layer, which predicts the final output among the hidden layers that perform most of the computations required by our network. Neurons forming the layers are activation functions that establish clear input in a simple ANN cell (Haykin, 2008).

Although ANNs are not in a single topology, many topologies can be established due to the combination of elements that make up ANNs. Topologies differ from each other with their decisive features, such as the number of layers used in the network (single-layer, multilayer), learning algorithm (supervised, unsupervised, reinforced), learning rule, and communication direction (feedforward, feedback, competitor). However, the basis of all ANNs models is single-layer sensors (perceptrons). Figure 2 presents the sensor's image, based on processing multiple inputs in a single layer and producing a single output.

Figure 2

A Simple Artificial Neural Network Cell Structure



Considering the cell structure of the ANN, the input variables spread towards the net function with certain weights. Weights are the values corresponding to the connections that enable neurons to transfer data to each other. Weights show the effect of the relevant input on the cell and the value of the information coming to the cell. The stage determining the best weight cluster in the ANN is the training (learning) stage or network training (Haykin, 2008). While assigning the weight values completely randomly before the training, the aim is to obtain the weight values that will ensure reaching the correct output by changing the weight values during the training or learning process.

The difference between the output variables (expected value) given to the neural network during the training phase and the predicted (observed) output is called error. An iterative process determines the weights when the error reaches below the desired level (threshold). While reaching the threshold value, the weight change is the learning rate. Choosing a too small or too large learning rate may cause excessive learning or missing the most appropriate weights (Anderson & McNeill, 1992). While the determined "threshold" and "learning rate" values determine the training period of the neural network, net and activation functions affect the structure of the output.

The inputs are multiplied by the weight values to form a new input value within the net (addition, multiplication, maximum, minimum) functions. The weighted values can be converted to a net input value by summing or multiplying according to the determined function, or the net input value can be determined by selecting the minimum maximum values. The net inputs are processed in the activation function and directed to the information output layer. The activation function contributes to determining the output value with different calculations. These calculations frequently use linear, step, sigmoid, tangent hyperbolic, and corrected linear unit functions (Haykin, 2008). Each of the activation functions has its advantages and disadvantages. However, there are no definite rules in selecting these functions, such as determining the learning rate, threshold value, layer and number of neurons. In other words, there is no standard set of rules for establishing the topology of the ANN. Thus, researchers cannot claim that they have reached the most appropriate solution to the problem. They can only state that they offer acceptable solutions (Anderson & McNeill, 1992). Researchers can determine the most appropriate model with the assessment metrics of the models. While assessment metrics allow comparison of models, they also provide information about model success.

One of the aims of this study is to compare the performance of ANN and HLM, for which PISA data has been utilised as a tool. ANN offers the capability to integrate numerous factors that influence the dependent variable due to its ability to model complex relationships. Consequently, while predicting outputs with ANN, we aimed to highlight the differences between this approach and the more traditional methodologies of HLM. To evaluate the performance of the models, various metrics were employed, including Mean Error, Mean Absolute Error, Relative Absolute Error, Root Mean Square Error, and Variance Explained. These metrics allowed for a comprehensive analysis of the accuracy and explanatory power of both models.

METHOD

Research Sample

The sample of this study consisted of 15-year-old students from Finland, Kosovo and Türkiye who participated in PISA in 2018. OECD used a stratified random sampling method for PISA studies. This method aimed to include subgroups in the sample according to their size in the universe. The study aimed to increase the variance in dependent and independent variables by selecting students with low, intermediate, and high reading literacy. Hence, the researchers selected 5649 students from Finland, 5058 from Kosovo, and 6890 from Türkiye. As a result of the data cleaning process, the researchers obtained a sample size of 15525 students consisting of 5007 students from Finland, 4053 students from Kosovo and 6465 students from Türkiye. The researchers examined the degree to which the variance in the dependent variable differed according to countries and in the whole data set with average and standard deviation values. Considering Table 1, the average of the dependent variable differed by country.

Table 1

Mean and Standard Deviation Values of Variables by Country and All Data Set

	Türkiye		Finland		Kosovo		All Data	
	$\bar{X}$	sd	$\bar{X}$	sd	$\bar{X}$	sd	$\bar{X}$	sd
PV_avg	467	82	531	89	360	63	460	103

Data and Dependent-Independent Variables

The study employed students' reading literacy scores from the PISA 2018 cognitive domain test as the predicted variable in HLM and ANNs analyses. Before starting the data analysis, the researchers determined seven variables obtained from the student level and two from the school level, which can affect the reading literacy of students. In determining the variables, missing data, feature selection method, and variables stated in the literature to affect reading literacy were effective.

To prepare for the analysis, the researchers first examined the missing data structure. Following Bennett's (2001) the guideline that analyses with more than 10% missing data are likely to be biased, cases and variables with substantial missing data were considered for exclusion from the dataset. To ensure reliable estimates in the multilevel analyses, the researchers considered the number of units at each level. Based on recommendations (e.g., Maas & Hox, 2005; McNeish & Stapleton, 2016) in the literature, schools with fewer than 20 students were excluded to maintain a balance between sample size and model complexity. After these processes, the missing data rates in the data set were below 10% for all three samples. The researchers performed Little's MCAR test by country to examine the randomness of the distribution of the available missing data. They determined that the missing data were completely randomly distributed in the three countries' data sets ($p \geq 0.05$). The researchers preferred the random forest missing data algorithm, known for its high prediction accuracy in data mining, to address the missing data. This algorithm has advantages such as being used at different scale levels, handling non-linear variables, and accounting for interaction effects (Tang & Ishwaran, 2017). The missing data were assigned separately according to the countries, considering their variance in the variables.

In another step of the data cleaning process, the researchers performed the extreme value analysis on a country basis. While examining whether the ratios of the number of students in the categories were appropriate for categorical variables, the researchers determined Mahalanobis distances in continuous data. Table 2 presents the data set after subtraction according to extreme values.

Table 2

Information on the Final Data Set

Country	Number of Schools	Number of Students	Ratio
Finland	198	5007	25.3
Kosovo	135	4053	30.0
Türkiye	179	6465	36.1

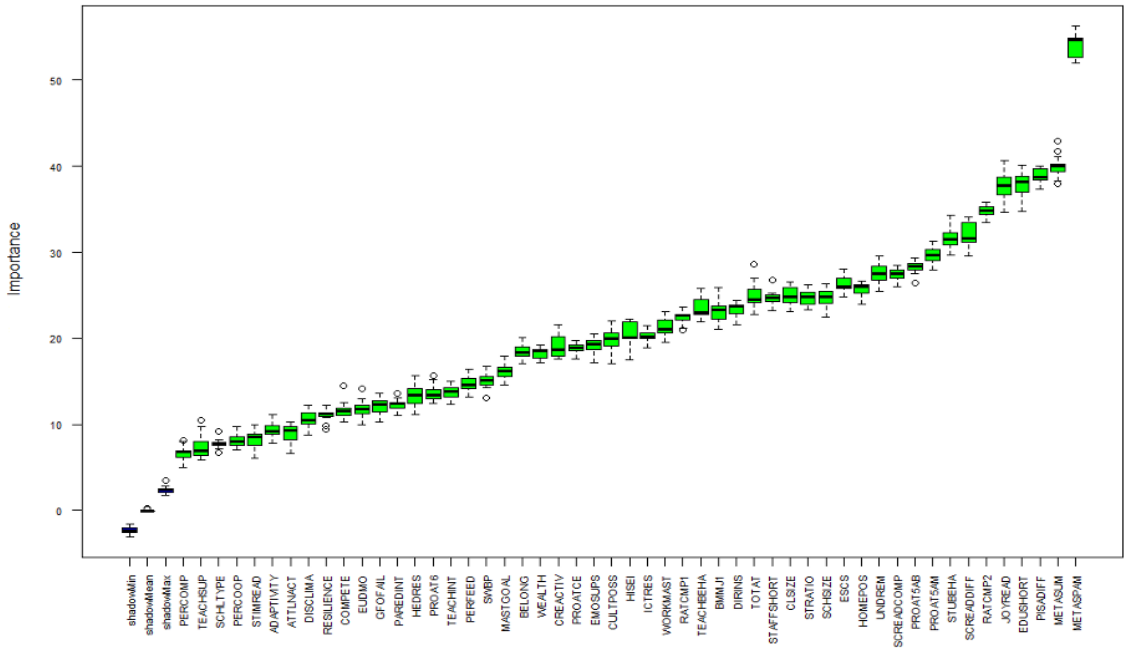
As Table 2, the researchers obtained a clean data set that met the necessary criteria (close sample sizes, more than 20 units at each level) for the planned analyses. In addition, they added the index variables of the schools to which the students in this data set belong and whose missing data were completed by the random forest method to the data set.

The feature selection method was used to choose predictive variables for this investigation. One justification for using this method was that the study compared HLM and ANNs analyses using the same variables as predictive variables. Another reason was that it was important to discover the optimum amount of model complexity in order to find the suitable point for variance and bias in the ANN data process (Munson & Caruana, 2009). This method is used to avoid the problem of overfitting, which means that the learning algorithm memorises the data. When too many variables are included, the model memorises the features and prediction accuracy may decrease. As a result, the Boruta feature selection algorithm was used in this study, which is one of the spiral approaches that successfully determines the best subfeature set. However, it should be noted that the random forest–focused feature selection approach of the Boruta algorithm may not fully accommodate HLM's sensitivity to multicollinearity nor meet ANNs' requirements for capturing complex interaction patterns; as such, it constitutes a methodological constraint that narrows the selected variable set and limits the generalizability of our model performance comparisons. Acknowledging this limitation, future assessments of alternative or complementary feature selection methods would enhance the study's depth.

The Boruta feature selection algorithm operates as an iterative process within the framework of random forests and has been employed for feature selection in this study. The fundamental principle behind the Boruta algorithm involves the use of shadow features, which are copies of the original features created by randomly shuffling the values within each feature. These shadow variables are then added to the dataset, resulting in an expanded dataset on which the random forest classification algorithm is trained. For each feature, the mean decrease accuracy metric is calculated to determine feature importance. In this iterative process, the importance measures of shadow features and real features are compared at each iteration to identify the significant real features. In this study, the minimum, average, and maximum values of the shadow variables generated by the Boruta feature selection method, along with the importance levels of the features, are presented in Figure 3.

Figure 3

Significance Level of Variables According to Boruta Feature Selection Method



As in Figure 3, all variables were significant as a result of the Boruta feature selection analysis. However, since all variables would complicate the model and make it difficult to interpret, the researchers selected nine variables important in the literature and ranking high in feature selection as predictive variables (Hu & Wang, 2022; Koyuncu & Fırat, 2020; Lim & Jung, 2014).These variables were "shortage of educational material (EDUSHORT)" and "shortage of staff (STAFFSHORT)" at the school level and "economic, social and cultural situation (ESCS)", "understanding and remembering (UNDREM)", "summarising (METASUM)", "assessment of reliability (METASPAM)", "disciplinary climate (DISCLIMA) ", "teacher-directed instruction (DIRINS)" and "teacher stimulated to read (STIMREAD)" at student level. The researchers aimed to examine the cognitive and external factors that may affect reading literacy in selecting these variables. Therefore, while the researchers preferred cognitive factors to determine reading comprehension strategies, they focused on these variables to examine external factors arising from teacher-student interaction and the learning environment. Finally, the study also included the ESCS variable, which greatly impacts academic achievement in the literature.

Data Analysis

As the PISA 2018 data used in this study has a nested hierarchical structure, it was preferred to utilize data analysis techniques that do not disregard the assumption of inter-unit independence. This study was planned according to two levels: students were the first level (Level 1), and schools were the second level (Level 2), which also included the first level. Following the data structure, the researchers examined variables affecting reading literacy with HLM and ANNs.

The researchers also used "ggplot2" (Wickham, 2016) package for the arrangement of the missing data and visualisations. Additionally, the researchers performed "Boruta" (Kursa & Rudnicki, 2010) feature selection analysis, "mlbench" (Leisch & Dimitriadou, 2021), "caret" (Kuhn, 2022), "randomForest" (Liaw, & Wiener, 2002), "neuralnet" (Fritsch et al., 2019) for HLM and ANN analyses, using R packages. Moreover, they supported HLM results with the HLM 6.0 program (Michela, 2006), as some results could not be obtained using only R packages or the HLM program and performed cross-validation of the analyses.

The researchers compared the results obtained in terms of evaluation metrics including mean error (ME), mean absolute error (MAE), relative absolute error (RAE), root mean square error (RMSE) and explained variance rate

(R^2) to compare the results obtained by ANNs and HLM. In notations from Equation 1 to Equation 5, $\hat{y}_i$ refers to the predicted value while y_i to the observed value:

ME (Mean Error): The mean error between a calculation model's predicted value and the observed value.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{1}$$

ME indicates the average bias of the predictions. A positive ME suggests overestimation, while a negative ME suggests underestimation. Ideally, ME should be close to zero. Values further from zero indicate systematic bias in the model.

MAE (Mean Absolute Error): The mean of the absolute values of the differences between the expected and observed values. It represents the mean amount of error in a prediction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

RAE (Relative Absolute Error): The ratio that compares a mean error with (residual) errors produced by an insignificant or pure model. An acceptable model will result in a ratio of less than one and close to zero.

$$RAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|\bar{y} - \hat{y}_i|} \tag{3}$$

RMSE (Root Mean Square Error): The fit index that shows the degree to which the model parameters are compatible with the universe covariances.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4}$$

R^2 (Variance Explained): It shows the extent to which the dependent variable (output value) is explained by the model's independent variables (input values).

$$R^2 = \left[\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)^2 \sum_{i=1}^{n}(y_i - \bar{y}_i)^2}}\right]^2 \tag{5}$$

Each of the metrics used in the study dealt with prediction errors from a different perspective and allowed us to evaluate the model regarding the direction, amount and distribution of errors. It's important to consider multiple metrics as each provides different insights. A model with lower error metrics (ME, MAE, RAE, RMSE) and higher R^2 is generally considered better. In this way, it provided a holistic perspective on evaluating the models (Botchkarev, 2019).

FINDINGS

The study employed models with similar structures as much as possible to compare the predictive performance of the ANNs method with HLM. In addition, the study also used the standardised data with an average of 0 (zero) standard deviation 1, which was created to establish the ANNs model for comparison purposes in the HLM model. The researchers compared and interpreted the performance of both models. Finally, the researchers examined the

capacity of the independent variables to estimate reading literacy specifically using Hierarchical Linear Modelling (HLM), the analysis method with high predictive performance.

Findings Regarding the Establishment of Data Sets

The researchers partitioned the data set for the ANNs model into a "training data set" comprising 75% of the observations, and a "test data set" containing the remaining 25%. This allocation enabled the neural network to be trained on 75% of the data, while reserving the remaining 25% for evaluating the model's generalisation capacity on unseen instances. As Hastie et al. (2009) point out, this data splitting strategy helps to mitigate overfitting and provides a more reliable estimate of the model's performance in practical applications.

While the researchers included all independent variables as input variables in the model during the establishment of the model in the training data set, they determined the average of the 10 plausible values (PV) in PISA 2018 as the output value. In the model determination process, the researchers preferred the most appropriate model, examining the models in different learning rates, threshold values, layer and neuron structures, and considering the learning time and evaluation metrics. The researchers used a three-layer topology consisting of three neurons for the model. The threshold value was 0.10, the learning rate was 0.2, and the linear function was determined as the activation function for all neurons. These values provided the best trade-off between convergence speed and stability of the learning process. Changing these values would indeed affect the model's performances. To evaluate the established model, the researchers ran the model again with the test data set. Table 3 presents the metrics obtained from the training and test data set.

Table 3

Artificial Neural Network Error and Explained Variance Results for Training and Test Sets

Metrics	Training dataset	Test dataset
ME	0.000	-0.007
MAE	0.546	0.538
RAE	0.656	0.651
RMSE	0.691	0.681
R^2	0.525	0.530

While there was no specific cut-off value for the metrics of the model established for prediction, the value R^2 was expected to approach 1, and the error values were requested to approach 0. Additionally, the similarity between the metrics calculated on the training and test datasets suggested the trained model had learned generalizable patterns rather than overfitting to the training data. As Hastie et al. (2009) discussed, achieving a balance between training and test errors, while keeping the errors reasonably low, enhances the model's generalization capability and allows for a more accurate assessment of its performance. Notably, if a model trained with very low training errors produces high test errors, it is an indicator of memorization rather than true learning. However, the close values between the training and test metrics reported in Table 3 indicated the established model was able to capture the relationships between the variables without overfitting.

Findings Regarding the Explained Variance and Error Values Obtained from the Analyses

To compare the performances of ANNs and HLM analyses, the researchers established two models on the same data set with the same dependent (output value) and independent (input) variables. Since it best fits the data in HLM analysis, the researchers preferred the intercept and slope model as the dependent variable. Additionally, the researchers used the explained variance and error values to evaluate the models. Since the values obtained from these evaluation metrics would provide information about the predictive performance of ANNs and HLM, the researchers compared the evaluation metrics, as in Table 4.

Table 4

HLM and ANNs Error and Explained Variance Results

Metrics	ANNs	HLM
ME	-0.002	-0.000
MAE	0.544	0.397
RAE	0.655	0.477
RMSE	0.689	0.506
R^2	0.526	0.744

Considering Table 4, the ME value in both methods was quite close to zero. While a ME close to zero can suggest that the established models did not make systematically biased predictions above or below the observed values (Chai & Draxler, 2014), it's important to note that this alone doesn't guarantee unbiased predictions. As Willmott and Matsuura (2005) point out, ME can be non-zero in regression models, and even when ME is zero, it doesn't necessarily indicate error-free predictions. Rather, it suggests that positive and negative errors may be balancing each other out. Therefore, while the near-zero ME values in our analyses are encouraging, they should be interpreted cautiously and in conjunction with other performance metrics to get a comprehensive understanding of model performance. Since the absolute value of the errors was not taken at this value, the fact that the potential value was zero did not indicate that predictions were error-free but that predictions were biased in a positive or negative direction.

The MAE metric, second metric in Table 4, represented the average prediction error per student (observed score - predicted score). The MAE was 0.544 for the ANN model and 0.397 for the HLM model. Given that standardised (z-point) versions of the mean PV scores were used in the analysis, this indicates that the mean prediction error for each student was approximately 50% of a standard deviation error in the ANN model and approximately 40% of a standard deviation error in the HLM model. The MAE metric clarified that HLM predicts with a lower mean absolute error than ANNs.

RAE, another evaluation metric that provides information about the errors, was 0.655 for the ANN and 0.477 for the HLM prediction. This value was equal to the ratio of the prediction error of the established model to the error of the primitive prediction model. Primitive prediction model can be considered as the simplest statistical prediction model. The obtained values showed that the error amount that would have been obtained if prediction was made with the primitive prediction model would be approximately 1.5 times (1/0 .655) the error obtained with ANNs and approximately 2.1 times (1/0 .478) the error obtained with HLM.

Since RMSE squares the difference between observed and predicted values, it is more negatively affected by large prediction errors. In this respect, RMSE provides information about the distribution of errors and the amount of errors. Theoretically, since the RMSE value can take a value from 0 to infinity, the 0.689 obtained from the ANNs model can only be significant compared to different models. In the HLM model, the RMSE metric was 0.506. R^2 the last evaluation metrics examined in this study were 0.526 in ANNs and 0.744 in HLM. It was quite high in HLM compared to ANNs, which represented the ability of both models to predict the observed values (R^2). Therefore, HLM showed better predictive performance in this data set than ANNs when using the same dependent and independent variables.

With the established ANNs and HLM models, the researchers predicted each student's score (PV values). The researchers established a scatter plot of the difference between each student's predicted and observed scores to examine how different the predicted scores were from the observed scores. Figure 4 presents the scatter plot of the error values of ANNs, and Figure 5 of HLM. Regarding the interpretability of the graphs, the error values were coloured according to the countries. The blue indicates Türkiye's error values, the orange indicates Finland's error values, and the green indicates Kosovo's error values.

Figure 4

Scatterplot of Error Values in ANNs

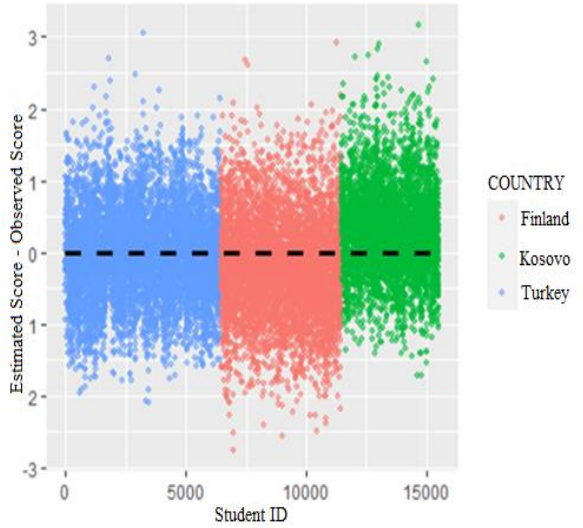
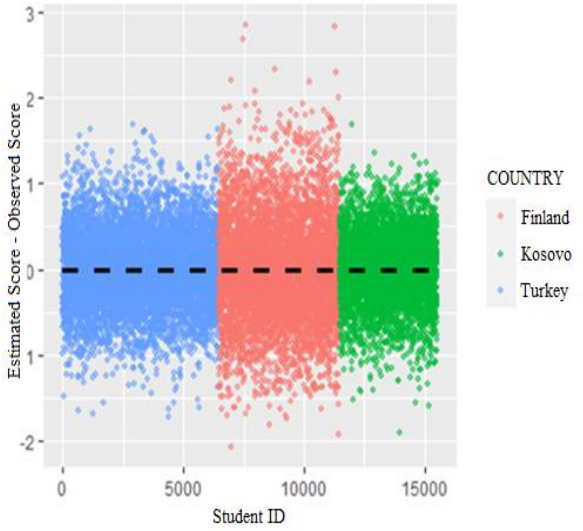


Figure 5

Scatterplot of Error Values in HLM

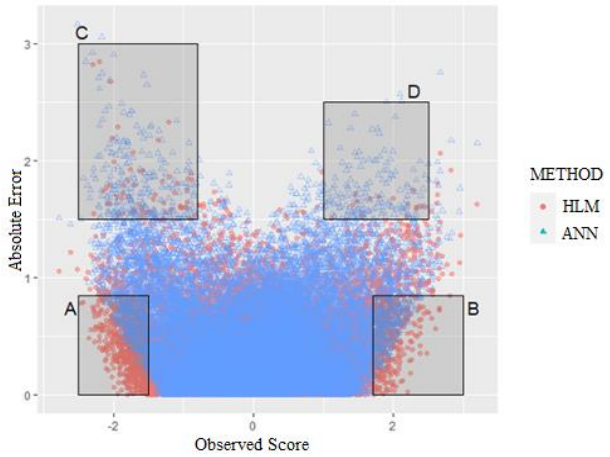


Considering the values of the ANNs model in Figure 4, the error values of Türkiye were distributed almost symmetrically around zero. In the error values of Finland, the negative error bias was higher than the positive error bias. On the other hand, considering the error values of Kosovo, the majority of the participants had positive error values, unlike Finland. Considering the values of the HLM model in Figure 5, the error values of students in all countries were almost symmetrically distributed around zero.

Moreover, the researchers established an absolute error scatter plot for ANNs and HLM to examine whether the absolute error values obtained from ANNs and HLM models differed in various regions of the observed score range. Figure 6 presents the graph established. In the scatter plot, the errors of the ANN model are represented in blue, and the errors of the HLM model are represented in orange.

Figure 6

Scatterplot of Absolute Error in ANNs and HLM



Considering the error values in Figure 6, the error values differed for ANNs and HLM at the ends of the observed score range. Boxes A and C or boxes B and D emphasised that the error values of HLM were concentrated in the range of 0 and 1 around the +2 and -2 values of the observed score, and the error values of ANNs were distributed in the range of 1.5 and 3. Therefore, according to boxes A, B, C and D, the errors obtained from HLM were less at extreme points than ANNs.

Findings Regarding the Reading Literacy Level of the Predictive Variables of Students in Finland, Kosovo and Türkiye

The study utilised all plausible values (PVs) for reading literacy to accurately estimate student performance in Finland, Kosovo, and Türkiye. Additionally, appropriate sampling weights were applied in the analysis to ensure the results are representative of the student populations in these countries.

The researchers established the random effects ANOVA (Model1) model below to examine whether there was a difference between the average reading literacy of the schools in Türkiye, Finland and Kosovo participating in PISA 2018. Table 5 presents the results of the Model1.

Table 5

Random Effects ANOVA Model Results

Fixed Effects		Coefficient	se	t ratio	p value
γ_{00}		461.281	3.693	124.896	0.000
Random Effects	Sd	σ^2	df	χ^2	p value
u_{0j}	82.233	6762.315	511	21514.7	0.000
r_{ij}	69.202	4788.950			

Considering Table 5, the average reading literacy score was 461, and the standard error was 3.693. Therefore, the standard deviation of reading literacy scores was 82.233, and the value ranges observed with a 95% probability of school averages were between 300 and 622 [461± (1.96x82.233)]. Besides, there were large differences between the average reading literacy scores of the schools in the sample. Also, the difference in average reading literacy scores between schools was statistically significant (χ^2= 21514.7; df=511; p< .001). When the interclass correlation of this difference was [6762/ (6762 +4789)= 0.59], 59% of the variance in reading literacy scores of students was in interschools.

In the second stage, the researchers established the averages model (Model2) as the dependent variable to determine which school characteristics affected the difference in students' reading achievement. In the third stage,

the researchers examined the random coefficient model (Model3) results to determine which student characteristics effectively differed in students' reading literacy. In the last stage, the researchers established an intersection and slope model (Model4) as a dependent variable to determine which school characteristics affected student characteristics that also affected students' mathematics achievement. Additionally, the researchers examined the deviation values of the models to determine the model that better fitted the data among the four models established. Since ten different plausible values were used as dependent variables, ten different deviation values were obtained for each model. Since it had the lowest deviation value for all plausible values, the model that best fitted the data was the intercepts and slope model as the dependent variable. Therefore, although the researchers established and examined all models, only the intersection and slope model results were given as dependent variables in the findings. Table 6 presents the results of this model.

Table 6

Results of Intercepts and Slopes Model as Dependent Variable

Fixed Effects		Coefficient	se	t ratio	p value
INTERSECT1, β_0					
INTERSECT2, γ_{00}		461.335	3.351	137.673	0.000
EDUSHORT, γ_{01}		-34.655	3.262	-10.625	0.000
STAFFSHORT, γ_{02}		10.112	3.740	2.704	0.007
ESCS, β_1					
INTERSECT2, γ_{10}		10.400	0.810	12.834	0.000
EDUSHORT, γ_{11}		0.152	0.752	0.202	0.840
STAFFSHORT, γ_{12}		-0.470	0.826	0.569	0.569
UNDREM, β_2					
INTERSECT2, γ_{20}		10.765	0.605	17.787	0.000
METASUM, β_3					
INTERSECT2, γ_{30}		11.805	0.729	16.190	0.000
EDUSHORT, γ_{31}		0.114	0.662	0.173	0.863
STAFFSHORT, γ_{32}		-0.355	0.794	-0.447	0.655
METASPAM, β_4					
INTERSECT2, γ_{40}		17.994	0.899	20.010	0.000
EDUSHORT, γ_{41}		-2.193	0.886	-2.475	0.014
STAFFSHORT, γ_{42}		1.126	1.028	1.095	0.275
DISCLIMA, β_5					
INTERSECT2, γ_{50}		2.862	0.643	4.452	0.000
EDUSHORT, γ_{51}		-0.032	0.623	-0.052	0.959
STAFFSHORT, γ_{52}		0.585	0.745	-0.786	0.433
DIRINS, β_6					
INTERSECT2, γ_{60}		-3.119	0.644	-4.840	0.000
STIMREAD, β_7					
INTERSECT2, γ_{70}		4.265	0.641	6.653	0.000
EDUSHORT, γ_{71}		0.619	0.674	0.918	0.362
STAFFSHORT, γ_{72}		0.071	0.718	0.098	0.922
Random Effects	sd	σ^2	df	χ^2	p value
u_{0j}	74.667	5575.131	509	23918.6	0.000
ESCS u_{1j}	8.814	77.682	509	752.9	0.000
METASUM u_{3j}	7.208	51.958	509	703.7	0.000
METASPAM u_{4j}	14.103	198.904	509	980.2	0.000
DISCLIMA u_{5j}	5.755	33.116	509	627.9	0.000
STIMREAD u_{7j}	3.300	10.892	509	549.9	0.102
r_{ij}	59.009	3482.117			

Considering Model 4, the predictor level2 variable regarding the slope coefficients β_2j and β_6j was not added. In Model 3, the effect of these variables (UNDREM and DIRINS) on reading literacy scores was similar in interschools. Therefore, no slope variance would need to be explained by level2 variables.

According to Table 6, after STAFFSHORT was taken under control, the effect of the EDUSHORT variable on average reading literacy scores was -34.655. Besides, after EDUSHORT was taken under control, the effect of the STAFFSHORT variable on average reading literacy scores was 10.112, and these effects were statistically significant ($p<0.01$). Thus, a one-unit increase in the school's EDUSHORT level led to a 34.655-unit decrease in reading literacy scores, while a one-unit increase in the STAFFSHORT level led to a 10.112-unit increase. In addition to the variables, ESCS, UNDREM, METASUM, METASPAM, DISCLIMA, DIRINS, and STIMREAD variables added at level 1 continued to have a statistically significant effect on reading literacy scores ($p<0.001$). When level1 variables were constant except for DIRINS, a one-unit increase in the DIRINS variable led to a 3.119-unit decrease in reading literacy scores. For other level1 variables, when all variables except the relevant variables were taken under control, a one-unit increase in these variables led to an increase in reading literacy scores of 10.400, 10.765, 11.805, 17.994, 2.862 and 4.265 units, respectively.

The researchers examined to what extent level2 variables explained the effect of level1 variables whose predictive power was examined on reading literacy scores. After controlling only the STAFFSHORT variable, the effect of the EDUSHORT variable on the METASPAM-reading literacy scores slopes was (γ_{41}) -2.193 and statistically significant ($p<0.05$). There was no statistically significant predictor of the slope between EDUSHORT and STAFFSHORT reading literacy scores and other level1 variables ($p>0.05$).

According to the random effects results of the intersection and slope model as a dependent variable, the variance of the average reading literacy scores of the schools at the school level was 5575 after EDUSHORT and STAFFSHORT were taken under control ($p<0.001$). The researchers compared the predicted variance values in Model3 and Model4 [(6827-5575)/6827] for the average school reading literacy scores. Accordingly, EDUSHORT and STAFFSHORT explained approximately 18% of the inter-school variance in reading literacy scores. The researchers also compared Model3 and Model4 variances to examine how the level2 variables explained the slope's interschool variance. However, only the residual variances of the METASPAM variable were compared, as EDUSHORT significantly explained the METASPAM-reading comprehension skill scores slope in the fixed effects of Model 4 [201-199/201]. Accordingly, EDUSHORT explained approximately 1% of the inter-school variance in METASPAM-reading literacy scores slopes. After the level2 variables were taken under control, the variance to be explained remained in the slope established by all level1 variables except the STIMREAD variable with the reading literacy scores ($p<0.001$).

CONCLUSION, DISCUSSION, AND RECOMMENDATIONS

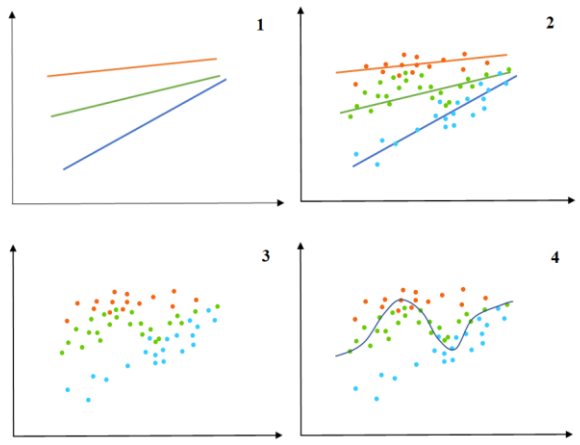
HLM is often one of the first analytical approaches considered when working with hierarchical data structures. However, the literature suggests employing alternative algorithms instead of analyses that focus on examining the influence of independent variables on dependent variables (Dong & Hu, 2019; Gamazo & Martínez-Abad, 2020; Oreta, 2004). Although these recommendations stem form research results aimed at comparing the analyses, some research results also suggest using traditional analyses (Feng & Jones, 2015). In addition, it is stated that ANNs are a powerful method for solving nonlinear problems (Kumova & Kışla, 2020). Consequently, this study sought to investigate the performance of HLM and ANNs in examining the variables impacting reading literacy, with the researchers aiming to utilize a highly performant method.

For these purposes, the researchers compared HLM and ANNs analyses regarding error metrics and explained variance. Furthermore, the researchers determined that all error metrics of HLM were lower than the error metrics of ANNs. Therefore, HLM analysis could make predictions with fewer errors in the data set. In addition, HLM had a higher explained variance value, which represented the relationship between the observed scores, which were evaluated as reading literacy, and the predicted scores. Although different results have been obtained from comparative studies such as this study, the interpretability of the results of the HLM has been emphasized (Levy et al, 2020; Weng & Luo, 2023). At the beginning of the study, the researchers predicted that the situation to be examined due to the multilevel data set may belong to a nonlinear problem. This opinion was based on the differentiation of the relationship between the first-level predictive variables and the predicted variable (output variable) according to the values of the second-level or higher-level variables in the hierarchical data. Considering

the fixed effects coefficients of the random coefficient model established when HLM analysis is carried out and the intersection and slopes model as dependent variables, the values differed. Therefore, regarding the two-level contextual model, the predictive variables added to the second level changed the slope of the line belonging to the group under consideration. The researchers considered that determining the curve of the observations instead of the predictive variable to be added to the second level and the lines formed with different slopes of the same group could establish a more effective prediction model. Figure 7 visualises the default state with four different graphs.

Figure 7

Observation Chart of Hypothetical Groups



In the first graph of Figure 7, lines show the relationship between the predicted and predictive variables of three hypothetical groups. The second graph includes the scattering of the variables represented by these lines. In HLM, the predictive power of the predictive variables is represented by the slope coefficients of the regression lines of the groups. However, considering the third graph, including only the scattering of the variables and excluding the regression lines, the relationship between the predicted and predictive variables of all three groups also indicated a nonlinear problem. In other words, considering the data of the three groups separately, there was a linear relationship between the variables. Still, considering the observations of all groups, a nonlinear relationship may occur. To visualise this situation, the researchers established a fourth graph. Therefore, if the hypothetical situation in Figure 7 occurred, the preference of multilayered ANNs for solving the nonlinear problem may contribute to establishing a more successful model. However, although multilayer ANNs had a high performance in solving nonlinear problems or showed better or similar performance compared to different regression models (Lin et al., 2012), ANNs could not show similar success in this multilevel data set. Table 4 presents the metrics for this result.

The error values of the analyses were examined by country and the entire data set. The researchers compared the error scatter plots showing the difference between the predicted and observed scores of the ANNs and HLM models established in Figures 4 and 5. As a result, in the ANNs analysis, the error values of Türkiye were almost symmetrically distributed around zero, the negative error bias was higher in the error values of Finland, and the positive error bias was in the majority in the error values of Kosovo. In HLM, the error values of students for all three countries were almost symmetrically distributed around zero. While Türkiye had an intermediate reading literacy level, Kosovo had a low reading literacy level. Finland had a high level of reading literacy. As the data moved away from the average value of the observed scores in both directions, the amount of error in ANNs increased, and HLM was unaffected by this situation. In other words, ANNs made relatively higher erroneous predictions at the extremes of the observed score gap of students. In addition, as seen in Figure 6, the researchers examined absolute error values in the entire data set. Then, they determined that the error values differed for ANNs and HLM at the extremes of the observed score range. The researchers also observed that the errors obtained from HLM were less at extreme points than ANNs. In other words, the fact that ANNs moved away from the score prediction observed at extreme points caused it to show a lower performance in the data set used from HLM. Therefore, ANNs may be similar to or higher than HLM in the data set formed by observations around the centre. For this reason, different researchers can further examine the effect of extreme data on the prediction process of

ANNs. Predictive performances of HLM and ANN can be examined separately in different distribution types (kurtosis, patency), and plausible critical values for this effect can be revealed.

Another reason for obtaining these results can be considered as the high correlation interclasses. In other words, the organisations at the first level were homogeneously distributed among the groups in which they were located at the higher levels. Therefore, examining this situation with simulation studies by producing data with different inter-class correlations can be presented to researchers as a suggestion. In addition, since the predictive level of the high-level variables affected the slopes of the regression lines of the groups, the predictive levels of the high-level variables can be artificially changed, and the researchers can examine the results of this effect in the future.

In comparing HLM and ANNs, the current research examined the relative importance of predictive variables on the predicted variable. The researchers observed that the predictor variables had different order of importance in both analyses. HLM and ANNs follow different paths in the prediction process. While HLM performs correlation-based prediction between dependent and independent variables, iterative methods predict the weight of ANNs input variables on the output variable. For this reason, the predictive variables have different orders of importance in the two analyses.

Following the comparison of HLM and ANNs, the researchers examined reading literacy with HLM, which had a higher performance in terms of prediction accuracy. The researchers determined the necessity of applying HLM analysis with the established random effects ANOVA model. The researchers concluded that the difference in the average reading literacy scores in interschool was statistically significant. In addition, they observed that 59% of the variance in reading literacy scores of students was interschool. Therefore, the researchers established the averages model as the dependent variable, adding different levels of predictive variables in different models, the random coefficient model, and finally, the intersections and slopes model as the dependent variable. They preferred the last established model to express the predictive power of independent variables because the model that adapted better to the data was the intersection and slopes model as the dependent variable during the examinations of the deviation values of the models.

In line with the results obtained from the averages model and the random coefficient model as dependent variables, the researchers added variables to the intersection and slopes model as dependent variables. The coefficients obtained from this model showed the final effect of the variables, as in Table 6. According to Table 6, a one-unit increase in the EDUSHORT level of the school led to a 34.655-unit decrease in the reading literacy scores of the schools. Moreover, a one-unit increase in the STAFFSHORT level of the school led to a 10.112-unit increase in the reading literacy scores of the schools. With this significant effect, EDUSHORT and STAFFSHORT explained 18% of the inter-school variance. However, there was still a significant variance to be explained in the schools' average reading literacy scores. These variables referred to the shortage of staff (STAFFSHORT) and materials (EDUSHORT) in schools. The presence of educational material shortages, such as EDUSHORT textbooks, information technology equipment, library, laboratory, and physical infrastructure deficiencies, such as buildings, heating/cooling, lighting, and acoustic systems, negatively affected students' reading literacy scores. Different studies examining the reading literacy scores of countries such as Albania and Kazakhstan in PISA 2018 found that the shortage of material did not affect reading literacy scores (Muratkyzy, 2020; Shahini, 2021). The reason why the result of this study is different from the literature may be the diversity of schools. In other words, since the studies in the literature are based on a single country, the shortage of materials in schools is at a similar level. This study included the countries with high, intermediate and low levels of reading literacy, and schools' material opportunities differed. The increase in the variance of material possibilities may have revealed the effect of reading literacy on scores. In line with the results obtained from this study, it is recommended to develop education and economic policies to prevent the shortage of physical and educational materials in schools. Studies to be carried out in this direction will contribute to equal opportunities in education nationally and internationally.

The present study concluded that the shortage of staff (STAFFSHORT), exemplified as the shortage/inadequacy of instructors and auxiliary personnel, positively affected students' reading literacy scores. Various studies conducted to examine the cause of this situation found quite different results. While Shahini (2021) stated that the staff shortages of Albanian citizens participating in PISA 2018 negatively affected their reading literacy scores,

Muratkyzy (2020) concluded that the staff shortage of Kazakhstani students participating in PISA 2018 application did not statistically affect their reading literacy scores. Since academic success may decrease when there is a shortage and staff inadequacy, the STAFFSHORT variable's effect in fields such as science and mathematics has also been a matter of curiosity. In the literature, Courtney et al. (2022) investigated the effect of STAFFSHORT on mathematics and science skill scores in all countries participating in the applications from PISA 2009 to PISA 2018. According to this study, while shortage and inadequacy of staff harmed mathematics skill scores in 2012, there was no statistically significant effect in the other three applications. Considering the effect on science skill scores, there was no statistically significant difference in 2018, while the other three applications had a negative effect. Comparing the studies in the literature and this study, the results obtained were almost the opposite of those in the literature. This may be because of the different sampling methods used in the studies. On the other hand, in this study, the perception of school principals about inadequacy and shortage of staff may have been effective in schools with high and intermediate reading literacy scores.

The first-level variables thought to affect reading literacy scores were ESCS, UNDREM, METASUM, METASPAM, DISCLIMA, DIRINS and STIMREAD. While all variables had a statistically significant positive effect on reading literacy scores, the DIRINS variable had a negative effect. When all other variables were kept constant, a one-unit increase in the DIRINS variable led to a 3.119-unit decrease in reading literacy scores. Among other studies investigating the effect of the DIRINS variable, which refers to teacher-directed education, on reading literacy scores, Hu and Wang (2022) conducted a study in 29 countries, including Türkiye and Finland, and Koyuncu and Frat (2020) examined data from PISA 2018 in China, Türkiye, and Mexico and discovered that had a significant negative effect on reading literacy scores. The DIRINS variable measures the teacher's basic teaching skills, such as "setting goals, making summaries, checking whether students are learning, giving feedback on what students should learn" in the lesson. However, since this scale has reverse scoring, the result obtained will increase the reading literacy scores of the students by increasing these teaching skills. Therefore, regarding the information obtained from this study and the literature in general, teacher-directed education positively affected reading literacy scores. Based on this result, teachers need to present specific information about the previous lesson or the subject covered that day at the beginning or end of the lesson. In addition, students should be informed about the subject to be covered at the beginning of the course. In other words, students should be informed about the target. Again, during the lesson process, whether the students have learned by using different teaching methods and techniques should be constantly checked. Feedback should be given to the students on the issues determined to be deficient and the issues they need to learn should be emphasised.

One of the first-level variables examined in the study was the ESCS variable. This variable, which expresses the economic, social and cultural status of students, was established by considering the parents' profession, education and household goods. When all other variables were kept constant, a one-unit increase in the ESCS variable led to a 10.400-unit increase in reading literacy scores. Many studies in the literature examining the effect of the ESCS variable on reading literacy scores observed that it had a statistically significant positive effect (Hu & Wang, 2022; Lee & Wu, 2013). In addition, Dong and Hu (2019) stated in their study that the effect of Singaporean students participating in PISA 2015 on their reading literacy scores was the most important variable among the 49 variables examined. Considering the results of the study and the literature, providing lifelong learning and adult education to improve families' economic, social and cultural situation should be among education's main objectives. It is necessary to raise students in the field needed to develop the economic situation of families and, therefore, to plan distant goals to ensure the balance of demand and supply in education. In addition, economic development is important in providing materials (internet, communication tools, art and education books, etc.) that will contribute to cultural development or provide the conditions and environment. For cultural development, Erasmus projects should be increased in schools, and development projects should be intensified, especially in rural areas. In addition, the accessibility of all students to cultural activities carried out by the Ministry of Culture and Tourism should be ensured for the participation of families in social activities.

Metacognition strategies knowledge was another concept that affected reading literacy scores in the study. UNDREM, METASUM and METASPAM variables are under the umbrella of metacognition strategies. UNDREM refers to the ability of students to understand and remember what they read, METASUM to summarise the text, and METASPAM to evaluate the content and reliability of the information in the text. When all variables except the related variables were controlled consecutively, a one-unit increase in UNDREM, METASUM and

METASPAM variables increased reading literacy scores by 10.765, 11.805 and 17.994, respectively. The study conducted by Koyuncu and Fırat (2020), which examined the effects of UNDREM, METASUM and METASPAM variables together, determined that it positively affected the reading literacy scores of Turkish, Chinese and Mexican students. Chen, Lin, and Chen (2021) examined the effect of all metacognitive strategies on the reading literacy scores of countries belonging to Eastern culture (Hong Kong, Japan, Macau, Singapore, etc.) and Western culture (Australia, Denmark, Finland, Germany, etc.) in PISA 2018 and concluded that they had a statistically significant and positive effect. According to the results obtained from the literature and this study, using metacognitive strategies is important in increasing reading literacy. Thus, it is necessary to encourage teaching effective metacognitive reading strategies to students. Students should be carefully observed in each reading task. To teach reading strategies for a certain purpose, student level, text, and the problems students face in the reading process should be determined. It is also necessary to determine the appropriate reading strategy for the student. The basis for this is that teachers decide which strategy to use best and in which situations. When literacy is taught and developed in primary school, classroom teachers have great responsibilities in increasing their reading literacy scores. Therefore, practical training should be included in the education of classroom teachers to identify and use appropriate reading comprehension strategies for students.

Two variables affecting the reading literacy scores based on the teacher-student relationship were added to the first-level model. One of these variables was DISCLIMA, which refers to the disciplinary climate in the classroom. The last variable whose effect the study examined was STIMREAD. This variable refers to the teacher stimulating his/her students to read and improve their reading literacy. In the study, a one-unit increase in the DISCLIMA and STIMREAD variables led to an increase of 2.862 and 4.265 units in reading literacy scores, respectively. Although some studies examined both variables separately in the literature (Chen et al., 2019; Ma et al., 2022), others examined the effect of both variables on reading literacy scores at the same time (Koyuncu and Fırat, 2020; Lim, & Jung, 2014; Ning et al., 2013). Chen et al. (2019) stated that there was a positive and significant relationship in their study examining the effect of teacher support and Ma et al. 's study investigating the effect of disciplinary climate in the classroom. The studies conducted by Koyuncu and Fırat (2020) and Ning et al. (2013) concluded that both the disciplinary climate in the classroom and teacher support positively affected the reading literacy scores. However, Lim and Jung (2014), who examined the effect of PISA 2009 on the reading literacy scores of Korean students, stated that only the disciplinary climate variable in the classroom had a positive effect, and teacher support had no effect. Concerning the results obtained from this study and the studies in the literature, the teacher-student relationship is important in increasing the reading literacy scores. Therefore, to increase the disciplinary climate in the classroom and support teachers in improving the student's reading by allocating more time to each student, the classroom size should be reduced, or an assistant instructor should be present in the classroom.

The findings of this study offer valuable implications, particularly for optimising educational practices and enhancing learning conditions in schools. Thus, it is suggested that these results could inform educational policies and practices. The performance differences between ANN and HLM underscore the importance of selecting predictive models that best align with the structure of the data and desired outcomes of educational assessments.

From a policy perspective, these results indicate that staffing levels and educational resources significantly affect variations in student performance. For example, the negative impact of educational resource shortages (EDUSHORT) on reading literacy underscores the need for policy interventions aimed at addressing disparities in resources. Similarly, the positive correlation between adequate staffing levels (STAFFSHORT) and reading literacy emphasises the potential benefits of investing in sufficient, qualified teaching staff.

In conclusion, such comparisons not only improve predictive accuracy but also support the development of evidence-based practices.

Acknowledgements and Explanations

No support was received from any institution, organization or person in this research. This study was produced from the doctoral dissertation conducted by the first author under the supervision of the second author.

Conflict of Interest

The researchers do not have any personal or financial conflicts of interest with other individuals or institutions related to the research.

Author Contribution

The first author contributed 70% and the second author 30% to this article.

REFERENCES

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon, 4*(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

Akkoyunlu, B. (2008, Mayıs). *Bilgi okuryazarlığı ve yaşam boyu öğrenme.* 8th International Educational Technology Conference (IECT) toplantısında sunulan bildiri, Anadolu Üniversitesi, Eskişehir.

Anderson, R.C., Hiebert, E.H., Scott, J.A. & Wilkinson, I.A.G. (1988). Becoming a nation of readers: the report of the commission on reading. *Education and Treatment of Children, 11*(4), 389-396.

Anderson, D., & McNeill, G. (1992). *Artificial neural networks technology*. Kaman Sciences Corporation

Ali, O., Murray, P. A., Momin, M., Dwivedi, Y. K., & Malik, T. (2023). The effects of artificial intelligence applications in educational settings: Challenges and strategies. *Technological Forecasting and Social Change, 199,* 123076. https://doi.org/10.1016/j.techfore.2023.123076

Baba, A. (2024). Neural networks from biological to artificial and vice versa. *Biosystems, 235*, 105110. https://doi.org/10.1016/j.biosystems.2023.105110

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health, 25*(5), 464-469.

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information Knowledge and Management, 14,* 045–076. https://doi.org/10.28945/4184

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chavez, H., Chavez-Arias, B., Contreras-Rosas, S., Alvarez-Rodríguez, J. M., & Raymundo, C. (2023). Artificial neural network model to predict student performance using nonpersonal information. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1106679

Chen, J., Lin, C. H., & Chen, G. (2021). A cross-cultural perspective on the relationships among social media use, self-regulated learning and adolescents' digital reading literacy. *Computers & Education, 175*, 104322.

Chen, Q., Lei, Y., Wen, Z., Li, S., Li, J., & Kong, Y. (2019). Teacher support, reading strategy and reading literacy: A two-level mediation model. *Best Evid Chin Edu, 2*(1), 157-170. https://doi.org/10.15354/bece.19.ar1036

Courtney, M., Karakus, M., Ersozlu, Z., & Nurumov, K. (2022). The influence of ICT use and related attitudes on students' math and science performance: Multilevel analyses of the last decade's PISA surveys. *Large-Scale Assessments in Education, 10*(1), 1-26. https://doi.org/10.1186/s40536-022-00128-6

Çelenk, S. (2003). Okul aile iş birliği ile okuduğunu anlama başarısı arasındaki ilişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 24*, 33-39.

Dong, X., & Hu, J. (2019). An exploration of impact factors influencing students' reading literacy in Singapore with machine learning approaches. *International Journal of English Linguistics, 9*(5), 52-65. https://doi.org/10.5539/ijel.v9n5p52

Ercan, U. (2021). Ev dışı gıda tüketim sınıflarının yapay sinir ağları ile tahmin edilmesi. *İşletme Araştırmaları Dergisi, 13*(4), 3265-3277.

Feng, Y., & Jones, K. (2015, July 8-10). *Comparing multilevel modelling and artificial neural networks in house price prediction* [Paper presentation]. 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services, Fuzhou, China.

Fritsch, S., Guenther, F., & Wright, M. N. (2019). neuralnet: Training of Neural Networks. R package version 1.44.2.

Gamazo, A., & Martínez-Abad, F. (2020). An exploration of factors linked to academic performance in PISA 2018 through data mining techniques. *Frontiers in Psychology, 11*, 575167. https://doi.org/10.3389/fpsyg.2020.575167

Guthrie, J. T., & Wigfield, A. (2000). Engagement and motivation in reading. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 403–422). Lawrence Erlbaum.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer Science & Business Media.

Hawkins, D. M. (2003). The problem of overfitting. *Journal of Chemical Information and Computer Sciences, 44*(1), 1–12. https://doi.org/10.1021/ci0342472

Haykin, S. (2008). *Neural networks: a comprehensive foundation.* New Jersey: Prentice-Hall.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2. ed): New York: Routledge.

Hu, J., & Wang, Y. (2022). Influence of students' perceptions of instruction quality on their digital reading performance in 29 OECD countries: A multilevel analysis. *Computers & Education, 189*, 104591. https://doi.org/10.1016/j.compedu.2022.104591

Kumova M., S. & Kışla, T. (2020). Yapay sinir ağları. In T. Güyer, H. Yurdugül & S. Yıldırım (Ed.), *Eğitsel veri madenciliği ve öğrenme algoritmaları* (pp. 127-146). Anı Publishing.

Koyuncu, İ., & Fırat, T. (2021). Investigating Reading Literacy in PISA 2018 assessment. *Lnternational Electronic Journal of Elementary Education*, *13*(2), 263–275. https://doi.org/10.26822/iejee.2021.189

Kuhn, M. (2022). caret: classification and regression training. R package version 6.0-91.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software, 36*(11), 1-13.

Leisch, F., & Dimitriadou, E. (2021). mlbench: Machine learning benchmark problems. R package version 2.1-3.

Levy, J., Mussack, D., Brunner, M., Keller, U., Cardoso-Leite, P., & Fischbach, A. (2020). Contrasting classical and machine learning approaches in the estimation of Value-Added scores in Large-Scale Educational data. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.02190

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18-22.

Lee, Y. H., & Wu, J. Y. (2013). The indirect effects of online social entertainment and information seeking activities on reading literacy. *Computers & Education, 67,* 168-177. https://doi.org/10.1016/j.compedu.2013.03.001

Lin, M. B., Groves, W. A., Freivalds, A., Lee, E. G., & Harper, M. (2011). Comparison of artificial neural network (ANN) and partial least squares (PLS) regression models for predicting respiratory ventilation: an exploratory study. *European Journal of Applied Physiology, 112*(5), 1603–1611. https://doi.org/10.1007/s00421-011-2118-6

Lim, H. J., & Jung, H. (2019). Factors related to digital reading achievement: A multi-level analysis using international large scale data. *Computers & Education*, *133*, 82–93. https://doi.org/10.1016/j.compedu.2019.01.007

Ma, L., Xiao, L., & Hau, K. T. (2022). Teacher feedback, disciplinary climate, student self-concept, and reading achievement: A multilevel moderated mediation model. *Learning and Instruction, 79*, 101602.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. Methodology: *European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

Michela, J. L. (2006). Software review: HLM 6. *Organizational Research Methods, 9(*1), 119-122. https://doi.org/10.1177/10944281052812

Munson, M.A., Caruana, R. (2009). On Feature Selection, Bias-Variance, and Bagging. Buntine. In W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (Eds.), *Machine Learning and Knowledge Discovery in Databases. ECML PKDD Lecture Notes in Computer Science, vol 5782*. Springer, Berlin, Heidelberg.

Muratkyzy, A. (2020). Equity and excellence in the Kazakhstani Education System: *A multilevel analysis of the personal and contextual factors contributing to students' reading literacy performance on PISA 2018* [Unpublished master's thesis]. Nazarbayev University Institute of Educational Sciences, Astana.

Ni, X. (2008). Research of data mining based on neural networks. *World Academy of Science, Engineering and Technology*, *39*, 381-384.

OECD. (2019). *PISA 2018 results volume I: What students know and can do.* Paris: OECD Publishing. https://doi.org/10.1787/5f07c754-en.

Oreta, A. W. C. (2004). Simulating size effect on shear strength of RC beams without stirrups using neural networks. *Engineering Structures, 26*(5), 681-691. https://doi.org/10.1016/j.engstruct.2004.01.009

Radoyevic, N. (2006). Exploring the use of effective learning strategies to increase students' reading comprehension and test taking skills (Yüksek lisans tezi). Brock Üniversitesi Eğitim Fakültesi, St. Catharines, Ontario.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rose, D. S., Parks, M., Androes, K., & McMahon, S. D. (2000). Imagery-based learning: Improving elementary students' reading comprehension with drama techniques. *The Journal of Educational Research, 94*(1), 55-63.

Shahini, A. (2021). Inequalities in Albanian education: Evidence from large-scale assessment studies. *Kultura i Edukacja, 4*(134), 40-70.

Snow, C. E. (2002). *Reading for Understanding: Toward an R&D Program in Reading Comprehension.* RAND Corporation.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining the ASA Data Science Journal, 10*(6), 363–377. https://doi.org/10.1002/sam.11348

Weng, W., & Luo, W. (2023). A comparative analysis of data mining methods and hierarchical linear modeling using PISA 2018 data. *International Journal of Database Management Systems*, *15*(2/3), 1–16. https://doi.org/10.5121/ijdms.2023.15301

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30, 79–82. https://doi.org/10.3354/cr030079

Yıldız Aybek, H. S., & Okur, M. R. (2018). *Predicting achievement with artificial neural networks: The case of Anadolu University Open Education System*. International Journal of Assessment Tools in Education, 5(3), 474-490.

Yousafzai, B. K., Khan, S. A., Rahman, T., Khan, I., Ullah, I., Ur Rehman, A., Baz, M., Hamam, H., & Cheikhrouhou, O. (2021). Student-Performulator: Student Academic Performance Using Hybrid Deep Neural Network. *Sustainability, 13*(17), 9775. https://doi.org/10.3390/su13179775

GENİŞLETİŞMİŞ ÖZET

Giriş

Bilgi edinme, işleme ve değerlendirme süreçlerinin teknolojik ve toplumsal gelişime katkısı arttıkça, okuma okuryazarlığı stratejik bir beceri olarak öne çıkmaktadır. PISA çerçevesinde okuma okuryazarlığı, bireylerin farklı biçimlerde sunulan metinleri anlama, kullanma, değerlendirme ve bu metinler üzerine düşünme yeterliği olarak tanımlanmaktadır. Bu çalışma, PISA 2018 uygulamasına katılan 15 yaş grubu öğrencilerinin okuma başarılarını etkileyen değişkenleri hem öğrenci (Düzey 1) hem de okul (Düzey 2) düzeylerinde incelemektedir. Analizlerde, hiyerarşik veri yapısına uygun Hiyerarşik Doğrusal Modelleme (HLM) ile bir veri madenciliği tekniği olan Yapay Sinir Ağlarının (YSA/ANN**)** yordama performansları karşılaştırılmıştır. Çalışmanın okuma başarısının belirleyicilerini çok düzeyli bir bakış açısıyla ortaya koymak ve aynı değişken seti kullanılarak HLM ve ANN yöntemlerinin tahmin gücünü karşılaştırmak olmak üzere iki temel amacı bulunmaktadır.

Yöntem

Bu çalışmada, evreni temsilen tabakalı rassal örnekleme yöntemiyle seçilen PISA 2018 verilerinden Finlandiya, Kosova ve Türkiye örnekleme dâhil edilmiştir. Veri temizliği ve kayıp veri işlemlerinin ardından örneklem, 15.525 öğrenci ve 500'ün üzerinde okuldan oluşmaktadır. Çok düzeyli analizlerde tahmin güvenirliğini artırmak amacıyla, 20'den az öğrencisi bulunan okullar veri setinden çıkarılmış; aykırı değerler ise ülke düzeyinde ayrı ayrı incelenmiştir. Kayıp verilerin dağılımı, ülke düzeyinde rastgelelik varsayımını karşılamış ve eksik veri ataması için doğrusal olmayan ilişkiler ve etkileşimlere duyarlı rastgele orman yaklaşımı kullanılmıştır.

Yordayıcı değişkenlerin belirlenmesinde, değişken sayısını yönetilebilir düzeyde tutmak ve aşırı uyum riskini azaltmak için Boruta özellik seçimi uygulanmıştır. Literatürde sıkça atıf alan ve önem sıralamasında öne çıkan dokuz değişken seçilmiştir. Bunlardan okul düzeyinde eğitim materyali yetersizliği (EDUSHORT) ve personel yetersizliği (STAFFSHORT); öğrenci düzeyinde ise sosyo-ekonomik-kültürel durum (ESCS), anlama-hatırlama (UNDREM), özetleme (METASUM), güvenirlik değerlendirme (METASPAM), sınıf disiplin iklimi (DISCLIMA), öğretmen-yönlendirmeli öğretim (DIRINS) ve öğretmenin okumaya teşviki (STIMREAD) değişkenleri analizlerde kullanılmıştır.

Hiyerarşik Doğrusal Modelleme (HLM) için klasik aşamalı modelleme yaklaşımı izlenmiş; rastgele etkiler varyans analizi ile çok düzeyliliğin gerekliliği sınanmıştır. Ardından sırasıyla ortalama modeli, rastgele katsayılar modeli ve kesişim-eğim modeli değerlendirilmiştir. Yapay Sinir Ağları (ANN) için ise veri %75 eğitim ve %25 test olarak bölünmüş, standartlaştırılmış girdilerle üç katmanlı bir ağ ayarlanmış hiperparametrelerle eğitilmiştir. Karşılaştırmada kullanılan ölçütler şunlardır: ortalama hata (ME), ortalama mutlak hata (MAE), göreli mutlak hata (RAE), kök ortalama kare hata (RMSE) ve açıklanan varyans (R^2). Bulgular, R paketleri ve HLM 6.0 yazılımı kullanılarak çapraz doğrulanmıştır. Yöntemsel sınırlılıklar kapsamında; Boruta'nın rastgele orman temelli yapısının, HLM'nin çoklu doğrusal bağlantıya duyarlılığının ve ANN'nin karmaşık etkileşim gereksinimlerinin tümüyle yansıtılamayabileceği dikkate alınmalıdır.

Bulgular

Aynı bağımlı ve bağımsız değişkenlerle kurulan iki model karşılaştırıldığında, Hiyerarşik Doğrusal Modelleme (HLM), Yapay Sinir Ağlarına (ANN) kıyasla tüm hata ölçütlerinde daha düşük değerlere ve daha yüksek R^2'ye ulaşmıştır. Hata yönünü gösteren ME her iki modelde de sıfıra yakın olmakla birlikte, genelleme başarısını yansıtan R^2'nin HLM lehine belirgin bir üstünlük sergilediği görülmüştür. Bu sonuçlar, söz konusu veri yapısında HLM'nin öngörü performansının daha güçlü olduğunu ortaya koymaktadır.

Öğrenci düzeyinde tahmin–gerçek farklarının dağılımı incelendiğinde, HLM'nin hata değerleri tüm ülkelerde sıfır çevresinde simetrik bir desen izlerken, ANN'de ülke bazlı yanlılık gözlenmiştir. Türkiye'de dağılım simetrik kalmış; Finlandiya'da negatif, Kosova'da ise pozitif yönlü bir sapma ortaya çıkmıştır. Ayrıca gözlenen puan aralığının uçlarında ANN'nin mutlak hata değerleri artış göstermiş, HLM ise uç gözlemlerde görece daha dayanıklı kalmıştır. Bu bulgular, ANN'nin örneklemin merkezinde HLM'ye yaklaşabildiğini, ancak uç gözlemler nedeniyle toplam performansının geride kaldığını düşündürmektedir.

ANN'nin eğitim ve test alt kümelerindeki hata ölçütlerinin birbirine oldukça yakın olması, aşırı uyumun gerçekleşmediğini ve modelin genel örüntüleri başarıyla öğrendiğini göstermektedir. Bununla birlikte, yüksek okullar arası benzerlik/dengelenmişlik (yüksek ICC) ve uç değerlerin varlığı, çok düzeyli yapıları açık biçimde modelleyen HLM karşısında ANN'nin avantajını sınırlamıştır. Kuramsal olarak, gruplar bazında doğrusal ilişkilerin tüm örneklem birleştirildiğinde doğrusal olmayan bir desene dönüşmesi durumunda çok katmanlı ağların üstünlük sağlayabileceği öngörülse de, bu çalışmada böyle bir durum gözlenmemiştir.

HLM'nin kesişim–eğim modeli sonuçlarına göre, okul düzeyi değişkenlerden EDUSHORT ile okuma puanı arasında anlamlı ve güçlü bir negatif ilişki bulunmuştur. Bu bulgu, okullarda materyal ve fiziksel altyapı yetersizliklerinin arttıkça öğrenci başarısının düştüğünü göstermektedir. Buna karşılık, STAFFSHORT değişkeni pozitif ilişki sergilemiştir. Literatürde bu değişkenin etkisine ilişkin bulgular tutarsızdır. Çalışmada gözlenen pozitif ilişkinin; (i) değişkenin okul müdürü algısına dayanması, (ii) ülkeler arası çeşitliliğin yüksek olması ve (iii) farklı örnekleme yaklaşımları gibi faktörlerden kaynaklanabileceği düşünülmektedir. İki okul düzeyi değişken birlikte ele alındığında, okullar arası varyansın önemli bir kısmını açıklamaktadır. Bununla birlikte anlamlı bir artık varyansın kalması, okul düzeyinde başka faktörlerin de etkili olduğuna işaret etmektedir. Ayrıca, EDUSHORT değişkeninin, bilginin güvenirliğini değerlendirme (METASPAM) stratejisi ile okuma başarısı arasındaki ilişkiyi zayıflattığı bulunmuştur. Bu bulgu, kaynak yetersizliğinin üstbilişsel stratejilerin getirilerini sınırlayabileceğini düşündürmektedir.

Öğrenci düzeyinde, ESCS (sosyo-ekonomik-kültürel durum), UNDREM (anlama–hatırlama), METASUM (özetleme), METASPAM (güvenirlik değerlendirme), DISCLIMA (sınıf disiplin iklimi) ve STIMREAD (öğretmenin okumaya teşviki) değişkenlerinin tamamı okuma başarısını pozitif ve anlamlı biçimde yordamıştır. DIRINS (öğretmen-yönlendirmeli öğretim) katsayısı ise sayısal olarak negatif görünmekle birlikte, ölçeğin ters puanlanmış olması dikkate alındığında sonuç, öğretmen-yönlendirmeli öğretimin okuma başarısını desteklediği biçiminde yorumlanmalıdır. Bu desen, hem bilişüstü stratejilerin öğretiminin hem de sınıf içi düzen ve öğretmen desteğinin öğrenci başarısına anlamlı katkı sağladığını ortaya koymaktadır.

Tartışma

Elde edilen bulgular, model-veri uyumunun önemini açık biçimde ortaya koymaktadır. HLM, hiyerarşik veri yapısını dikkate alarak düzeyler arası bağımlılığı açıklamakta; okul ve öğrenci düzeyi etkilerini ayrı ayrı ve etkileşimli biçimde modelleyebilmektedir. Bu özelliği sayesinde özellikle uç gözlemlerde daha istikrarlı tahminler üretmiştir. ANN'nin görece zayıf performansı ise, çok düzeyli yapıda yüksek grup etkisinin varlığı, uç değer duyarlılığı ve seçilen yordayıcıların etkileşim örüntülerini sınırlı biçimde temsil etmesiyle ilişkilendirilebilir. Bununla birlikte, ANN'nin eğitim verisine aşırı uyum göstermemesi, uygun ön-işleme ve model ayarlarıyla farklı senaryolarda daha rekabetçi olabileceğine işaret etmektedir.

Yöntemsel açıdan, özellik seçiminin rastgele orman temelli olması, HLM'nin çoklu doğrusal bağlantıya duyarlılığını ve ANN'nin karmaşık etkileşimleri yakalama kapasitesini eş zamanlı olarak en uygun biçimde desteklememiş olabilir. Bu nedenle, gelecekte LASSO, elastic net gibi karşılaştırmalı seçme–bedelleme yaklaşımlarının ya da çok düzeyli Boruta türevlerinin kullanılması önerilmektedir. Ayrıca, uç değerlerin ANN tahminlerine etkisinin incelenmesi, farklı okullar arası korelasyon düzeylerinin simülasyonu ve dağılım özelliklerine (basıklık/çarpıklık) göre ayrı değerlendirmeler yapılması, ilerideki çalışmalar için katkı sağlayacaktır.

Çalışma, PISA 2018 verilerinde üç ülke örneği üzerinden HLM'nin ANN'ye kıyasla üstün yordama performansı sergilediğini göstermiştir. Bu bulgu, eğitimde veri bilimi uygulamalarında veri yapısına uygun model seçiminin kritik rolünü teyit etmektedir. Okul düzeyinde, materyal yetersizliklerinin okuma başarısını düşürdüğü; personel yetersizliği göstergesinin ise bağlamsal nedenlerle beklenmedik biçimde pozitif yönde ilişkili olduğu belirlenmiştir. Öğrenci düzeyinde ise sosyo-ekonomik kaynaklar, bilişüstü okuma stratejileri, sınıf iklimi ve öğretmen desteği başarıyı artıran faktörler olarak öne çıkmaktadır. Ayrıca, öğretmen-yönlendirmeli öğretim değişkeninin ters puanlama yapısı dikkate alındığında, öğretmen rehberliğinin de öğrenci başarısına olumlu katkı sunduğu görülmektedir.

Sonuç olarak, çok düzeyli eğitim verilerinde HLM'nin açıklayıcılığı ve istikrarı, ANN gibi esnek yöntemlerle karşılaştırıldığında dahi yüksek kalmıştır. Bununla birlikte, uygun özellik seçimi, uç gözlemlerin yönetimi ve gruplar arası örüntülerin açık temsili sağlandığında, ANN'nin de belirli koşullarda güçlü bir alternatif oluşturabileceği unutulmamalıdır. Bu karşılaştırmalı yaklaşım, hem tahmin doğruluğunu artırmakta hem de kanıta dayalı eğitim politikalarının geliştirilmesi için sağlam bir zemin sunmaktadır.