Performance Prediction Data Mining System for Disabled Students Using Machine Learning

Kumar Pradyot Dubey[1], Narendra Kumar Gupta[2], Aditi Sharma*[3,4]

[1]Department of CS&IT, SHUATS, Prayagraj, UP, India
[2]Department of CS&IT, SHUATS, Prayagraj, UP, India
[3]Department of CSE, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India
[4]SMIEEE, SIT, Symbiosis International (Deemed University), Pune, India

Emails: pradyotdubey@gmail.com; narendra.gupta@shiats.edu.in; aditi.sharma@ieee.org

Abstract

This study aims to explore the educational achievements of individuals aged 21 to 38, specifically examining the differences between those with disabilities and those without. The research delves into the realm of Online Learning Platforms, which are recognized for offering extensive online courses that cater to both educational institutions and individual learners. Additionally, the study investigates Collaboration and Communication Platforms, which are designed to enhance interaction and cooperation among students and educators through various tools like discussion forums, chats, and shared workspaces. Adaptive Learning Platforms: Employing advanced algorithms and data analytics, this study used a dataset covering the UK from July 2013 to June 2020 to examine the highest skill levels of these two different groups. The data set, originally in Excel format, was carefully organized and structured for analytical purposes. The approach included the use of Python libraries such as NumPy for numerical calculations, and Matplotlib for visualization and proposed integration in a cloud-based system. The study's methodology is underpinned by sophisticated data analysis techniques, utilizing Python libraries such as NumPy, renowned for its efficiency in handling complex numerical calculations, and Matplotlib, which offers powerful visualization tools that are instrumental in elucidating the trends and patterns within the data. It is not only robust but also versatile, accommodating the integration of additional Python libraries such as Pandas for data manipulation and SciPy for more advanced scientific computations, thereby enhancing the depth and breadth of the analysis. Furthermore, the proposed integration of this analytical setup into a cloud-based system underscores the study's forward-thinking approach, aiming to leverage the scalability, accessibility, and collaborative potential of cloud computing. This integration promises to streamline the data analysis process, facilitating real-time data processing and enabling a dynamic exploration of the dataset. The study's methodology is underpinned by sophisticated data analysis techniques, utilizing Python libraries such as NumPy, renowned for its efficiency in handling complex numerical calculations, and Matplotlib, which offers powerful visualization tools that are instrumental in elucidating the trends and patterns within the data. This analytical framework is not only robust but also versatile, accommodating the integration of additional Python libraries such as Pandas for data manipulation and SciPy for more advanced scientific computations, thereby enhancing the depth and breadth of the analysis.

Keywords: Educational Attainment; Disability Status, Machine Learning; Data Visualization; Linear Regression; Numerical calculation; Cloud Based System; Educational platforms; Digital tools educational content; Digital Learning; Principal Component Analysis; Machine Learning.

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

1. Introduction

Educational institutions have embraced interactive learning tools such as game-based learning, virtual reality, and e-learning systems through learning management systems. This advancement has enabled the collection and analysis of student data for educational research. One study used Decision Tree, Neural Network, and SVM algorithms to assess how student internet behaviours can predict academic performance, achieving 71%-76% accuracy. Another research developed a machine learning-based system to classify students' academic performance as either good or bad, utilizing K-nearest neighbour and Decision tree classifiers. The Decision tree classifier showed a 94.44% accuracy rate in this context. Both studies primarily focused on accuracy as the metric for evaluating performance. Data was presented through various visualizations to show trends and patterns in educational performance over time.The research utilizes a comprehensive dataset, applying Python libraries such as NumPy and Matplotlib for data analysis and visualization, inspired by the systematic reviews and research findings in educational technology and machine learning, as outlined by Sekeroglu et al. [5] and Balaji et al. [2], among others. The integration of cloud computing is proposed to augment the accessibility and efficiency of educational data analysis, offering insightful perspectives for educators, policymakers, and researchers, echoing the extensive reviews by Balaji et al. [2] and Okoli et al. [28].

Data spanning from July 2013 to June 2020 within the UK was meticulously analysed to discern the highest educational achievements across both groups. Initially formatted in Excel, the data was restructured for in-depth analysis using Python libraries like NumPy for numerical analysis and Matplotlib for graphical representation. This analytical approach, drawing inspiration from the systematic reviews and empirical studies within educational technology and machine learning by Sekeroglu et al. [5] and Ofori et al. [10], highlights the transformative potential of digital platforms in advancing educational results, particularly for those with disabilities. The proposed integration with cloud computing seeks to expand the reach and efficiency of educational data analysis, thereby offering critical insights for educators, policy makers, and scholars, aligning with the extensive reviews by Balaji et al. [2] and Okoli et al. [28].The Growing Importance of Machine Learning in Education, Academic institutions face numerous challenges, including providing quality education, developing systems for assessing student performance, and identifying learners' future needs. The computerization of school data management has led to a paradigm shift, with an increased focus on analysing performance characteristics using academic and non-academic factors. This change has sparked interest in the machine learning community, leading to innovative approaches in educational data analysis. The two important aspects highlighted in this study are:

I. The Proportion of students with a 'Degree or Equivalent' Qualification: This aspect focuses on the percentage of individuals within the study's age group who have attained higher education qualifications, such as a bachelor's degree or an equivalent level of education. This metric is crucial for understanding the level of formal education achieved by the population and serves as a key indicator of educational attainment. This proportion allows for a nuanced understanding of the educational landscape, especially in terms of higher education access and completion rates among individuals with and without disabilities.

II. The Proportion ofstudents with 'No Qualification': This second aspect examines the percentage of individuals who have not obtained any formal educational qualifications. This measure is equally important as it sheds light on the lower end of the educational spectrum, highlighting those who may be at a disadvantage in the labour market and more broadly in society due to a lack of formal education. Understanding this proportion is vital for identifying groups that may require targeted educational support or intervention programs to improve their educational and subsequent employment outcomes.

A. Identified Gaps and Area of Improvement

There's a clear need for deeper understanding of how self-regulated learning strategies impact educational results, especially for people with disabilities, as underscored by prior studies from Issah et al. and Balaji et al. [1][2]. Moreover, there's a lack of thorough knowledge regarding the effectiveness of various educational platforms in fostering self-regulated learning outlined by Baashar et al. [3]. Although current studies, such as the one conducted by Issah et al. [1], provide important information, they also reveal significant areas that require further research. These platforms facilitate

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

structured educational content alongside support for autonomous learning, as investigated by Albreiki et al. [7] and Nawang et al. [6].These platforms underscore the significance of community in the learning process, as discussed by Makhtar and Nawang [6] and Issah et al. [1].

Advanced ML Techniques: Utilizing deep learning models can enhance the predictive accuracy and handle the complexity of educational datasets.

Do not cover diverse educational levels and neglect psychological and environmental factors that can impact performance.

Need for Advanced ML Techniques: The application of basic machine learning methods is prevalent, but there's a growing need to integrate more advanced techniques, like deep learning, to handle complex datasets and provide more accurate predictions.

Benchmark Datasets: Incorporating comprehensive datasets reflecting learners' performance from the beginning of their school lives can provide a more nuanced understanding of academic progress. Integration of Psychological and Environmental Factors: Including these factors could offer a more holistic view of student performance prediction.

Hands-on Data Modelling and Visualization: Moving beyond theoretical analysis to practical data modelling and visualization can provide more applied insights into educational outcomes.

Gaps in Research Due to Lack of Comprehensive Datasets and Basic ML Techniques:

Lack of Comprehensive Datasets: This point emphasizes the issue that many studies fail to encompass diverse educational levels and overlook crucial psychological and environmental factors that can significantly impact learning outcomes. This gap in research limits the depth and applicability of findings, particularly in understanding the full spectrum of factors influencing educational success among individuals with disabilities.

Need for Advanced ML Techniques: Here, the focus is on the prevalent use of basic machine learning methods in educational research. There's an articulated need for integrating more sophisticated approaches, like deep learning, to better handle the complexities inherent in educational datasets. This advancement could lead to more accurate predictions and insights into the learning process.

The Need for Integration of Holistic Factors and Practical Application in Research:

Integration of Psychological and Environmental Factors: This segment advocates for a more holistic approach to research by including psychological and environmental factors in the analysis. Recognizing these dimensions could lead to a more comprehensive understanding of student performance, especially among those with disabilities.

Hands-on Data Modelling and Visualization: The emphasis here is on moving beyond merely theoretical research to adopt more practical, hands-on approaches like data modelling and visualization. This shift could provide more tangible, applied insights into how educational strategies impact learning outcomes, thereby fostering a more effective educational environment for self-regulated learning.

B. Self-Regulated Learning

The field of machine learning's role in forecasting and improving student academic achievements is extensively documented and varied. Li et al. work [9] adds significant value to this area by methodically examining machine learning techniques for forecasting student outcomes. Contributions from Ofori et al. [10] and Enughwure et al. [11] further our collective grasp of the technological capabilities and constraints within the educational sector. In-depth analyses byBista et al. and Pandey et al. [13-15] explore the choice of algorithms and identify critical gaps in the existing literature, with Bista et al. [13] encouraging continued scholarly discussion on optimal practices and avenues for future inquiry. Comparative studies by Anuradha et al. [16] and methodological research by Sumitha et al. and Anuradha et al. [17, 18] provide frameworks for conducting systematic literature reviews, ensuring the thoroughness and dependability of research in this rapidly developing field. Targeted investigations [19-21] by Chair et al. and Lazzeroni et al. [19, 21], along with Sameer et al. [20], into various aspects of educational data mining showcase innovative strategies devised to tackle the challenges of evaluating and boosting academic success through data-centric approaches. This body

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

of work also underscores the intricacies and ethical considerations involved in deploying these technologies, laying a solid groundwork for subsequent research in this arena that improves students' performance and health in Pandey et al. [29], Winne et al. [30] and Reber et al. [32] research.

2. Benchmark Analysis Of Existing Methodology

Bressane et al. suggest the application of AI technologies to examine the relationship among study techniques, learning disabilities, and academic achievements. This approach utilizes Artificial Neural Networks (ANNs) to identify trends in the connection between these elements, followed by the implementation of Fuzzy logic-based AI for personalized educational strategies. A detailed benchmark study reveals that existing approaches often neglect to incorporate aspects of self-regulated learning and fail to fully leverage sophisticated machine learning algorithms for accurate forecasting of educational results by Nawang et al. and Albreiki et al. [6][7]. The study acknowledges that learning disabilities significantly affect academic performance, and traditional education models often fail to meet these students' unique needs. It underscores the importance of innovative, AI-driven methods in understanding and addressing these challenges. The proposal emphasizes AI's potential in educational settings, particularly for students with learning disabilities. AI tools can help educators and policymakers make data-driven decisions to enhance teaching methodologies and address the varying needs of all students, promoting inclusivity and equity in education.

This study responds to a critical research gap in the application of AI in education, especially concerning students with learning disabilities. While existing studies have shown the potential of AI in personalized learning, a detailed exploration of its role in optimizing educational approaches for students with learning disabilities is still needed. The research has significant practical implications. By adopting the proposed AI-driven strategies, educators can tailor their teaching methods to accommodate diverse student needs. This approach could revolutionize how educational institutions cater to students with learning disabilities, leading to improved academic outcomes and a more inclusive learning environment. Issah et al. [1] and Balaji et al. [2] set the groundwork by integrating diverse cognitive and non-cognitive factors into predictive models, illustrating the intricate nature of academic success. Comparative studies, such as those by Pandey and Taruna [15], assess different machine learning models, establishing benchmarks for future research in terms of method selection and application context. These benchmarks are critical for understanding the efficacy of different approaches and for guiding the selection of appropriate machine learning techniques to improve educational outcomes. The extensive review and analysis provided across these references serve as a valuable resource for researchers and educators aiming to utilize machine learning to enhance academic performance prediction and intervention strategies.

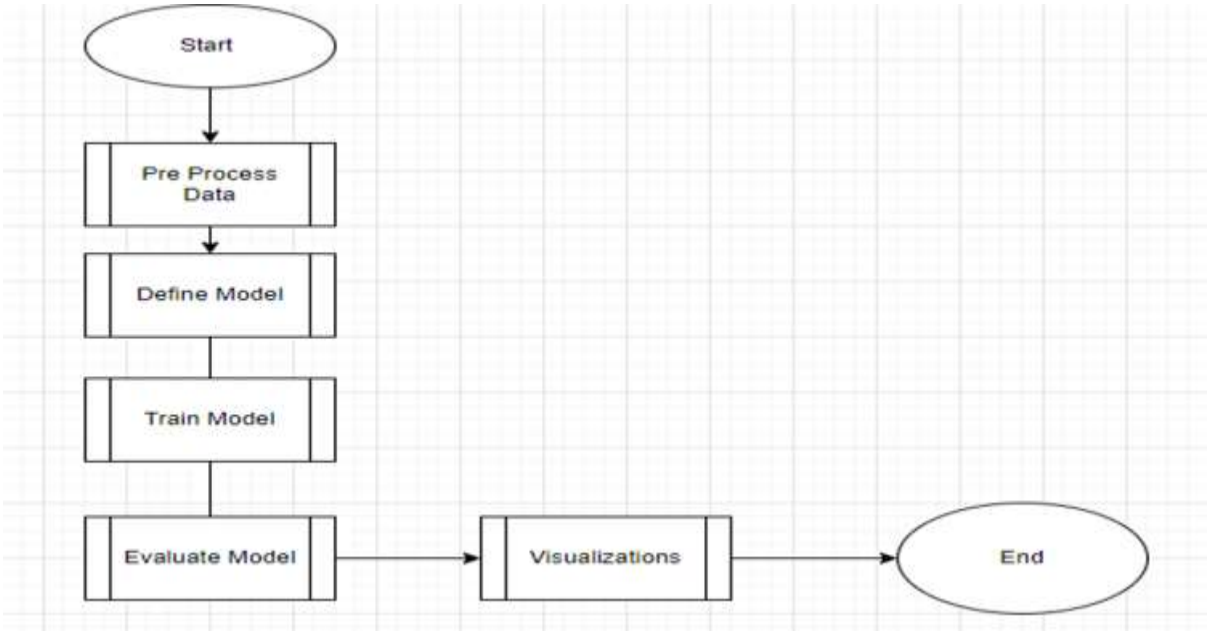

Figure 1: Proposed Methodology Flowchart

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

3. Proposed Framework

Here are the standard steps:

- When implementing linear regression in PyTorch, a powerful and flexible deep learning library, these principles are applied within a computational framework that facilitates both the definition of the model and the optimization process. PyTorch provides an intuitive interface for model construction, loss computation, and gradient descent optimization, making it an ideal choice for both beginners and experienced practitioners in the field of machine learning.

- Data standardization is a crucial preprocessing step in many machines learning pipelines, including linear regression. By transforming the features to have a mean of 0 and a standard deviation of 1, according to the formula $z = (x - \mu) / \sigma$, where μ is the mean and σ is the standard deviation, we ensure that the model treats all features equally, improving the convergence stability and speed during the training process.

1. Model initialization: An instance of the linear regression model is created.
2. Set the loss function and optimizer: The loss function (MSE) and optimizer (SGD) are set. The loss function measures the model's prediction error, and the optimizer updates the model's weights.
3. Epoch Loop: A loop is configured to loop over the training set multiple times (epochs) to train the model.
4. Zero gradient: To prevent accumulation, the gradient is reset to zero before each forward pass.
5. Forward pass: The model's predictions are calculated.
6. Loss Calculation: Loss is calculated by comparing the predicted and actual values.
7. Backward pass: gradients are computed for each model parameter.
8. Update weights: The optimizer updates the model weights based on the gradient.
9. Predicting test data: The model makes predictions on the test set.
10. The proposed framework aims to embed self-regulated learning components within educational platforms, utilizing machine learning models to tailor learning experiences and ameliorate educational results by Davis et al. [8].

A. Linear Regression:

Linear regression is a fundamental algorithm in the field of machine learning and statistics, used for predicting a quantitative response variable based on one or more predictor variables. It assumes a linear relationship between the input features and the target variable. This concept can be succinctly expressed by the simple equation $y = wx + b$, where y represents the predicted value, w is the weight coefficient that scales the input feature x, and b is the bias, an additional parameter that adjusts the output independently of the input value.

Step-I

Linear Regression using PyTorch-

Simple Equation: Linear Regression can be represented as

$y=wx+b$

y is the predicted value.

w represents the weight coefficient.

x is the input feature.

b is the bias

Data Standardization: Standardizing features to have a mean of 0 and a standard deviation of 1.

Simple Concept: $z= (x-\mu)/\sigma$

Step- II

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

Feature Selection Using Tree-based Models and RFE

Tree-based models like Decision Trees, Random Forests, and Gradient Boosting Trees can also be used for feature selection because these models provide a feature_importances_ attribute that ranks features based on their importance. The importance is typically calculated based on the reduction in node impurity (e.g., Gini impurity for classification, variance for regression) that each feature provides when used in the trees.

Here's how RFE and Tree-based models can be implemented for feature selection:

Train the Model: The estimator is trained on the initial set of features, and the importance of each feature is obtained either through the coef_ or feature_importances_ attribute.

Remove Features: The least important features are pruned from the current set of features. This number can be defined by the user or automatically determined based on the step parameter of RFE.

Repeat: Steps 1 and 2 are repeated with the reduced set of features until the desired number of features is reached.

RFE can be particularly powerful when combined with cross-validation via the RFECV class in scikit-learn, which performs RFE in a cross-validated loop to find the optimal number of features.

Both RFE and tree-based feature selection methods are powerful tools in the feature selection process, helping to improve model interpretability, reduce overfitting, and possibly enhance model performance by optimizing redundant features.

B. Training the machine learning model

Training a machine learning model involves adjusting parameters to minimize the error between predicted and actual results. This phase involves repeatedly running the training data through the model, calculating the loss, and adjusting the weights by back propagation. The machine learning model undergoes training utilizing a dataset encompassing variables pertinent to self-regulated learning, educational attainment, and other cognitive and non-cognitive factors, ensuring an exhaustive analysis by Enughwure et al. [11].

Epoch: An epoch represents a complete path through the training data set. Multiple epochs are often required to properly train a model.

Batch size: refers to the number of training samples used in one iteration. For simplicity, this example assumes that the batch size is the entire data set.

Training loop: Forward pass: Each training sample is passed to the model to obtain a prediction.

Loss calculation: Quantify the difference between predicted and actual values using a loss function.

Backward pass: The slope of the loss is calculated for the model parameters.

Optimization step: Model parameters are updated to reduce loss.

C. Correlation-Matrix

A heatmap of the correlation matrix can display the intensity and nature of linear associations among variables. Such a matrix is employed to identify the connections between various factors, shedding light on key determinants of educational success and the effects of self-regulated learning techniques by Gao et al. [12].

Group comparisons: These are particularly valuable for analysing these metrics across distinct groups.

Time series analysis: This is used to track long-term patterns, cyclical changes, and seasonal variations in the data.

Line chart: This is typically the most effective method for illustrating data over time.

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

D. Advanced Data Visualization and Dimensionality Reduction

High-dimensional data sets can make discerning relationships between variables challenging. By integrating sophisticated data visualization methods, we can more clearly and intuitively depict complex relationships in the data, aiding in a deeper understanding of the factors that impact educational achievements by Pandey et al. and Anuradha et al. [15][16]

E. Regression Equation

The formula y = mx + c represents a linear regression model where y is the forecasted percentage for individuals with disabilities, x is the percentage for those without disabilities, m denotes the regression line's slope (reflecting the variation in y with each unit increase in x), and c stands for the y-intercept. This regression formula, resulting from the linear regression analysis, measures the relationship between self-regulated learning techniques and educational achievement, providing a framework for predicting academic outcomes by Sumitha et al. [17].

4. RESULTS AND DISCUSSIONS

A. Experimental setup

nn.Module: At the heart of PyTorch's neural network module is the nn.Module class. It serves as the base class for all neural network modules, including layers, loss functions, or entire models. By subclassing nn.Module, you gain access to a wealth of built-in functionalities, including the ability to manage parameters, move your model to different devices, and apply built-in methods for training and inference.

__init__ Function: The __init__ function acts as the constructor for your model class. It's in this method that you define and initialize the various layers and components of your model. For a linear regression model, the initialization typically involves setting up a linear layer using nn.Linear. This layer encapsulates the weight and bias parameters, which the model learns during training.

nn.Linear: The nn.Linear class represents a linear transformation, which is the cornerstone of linear regression. It maps incoming data using a linear function defined by the equation y = wx + b, where w represents the weight and b the bias. For simple linear regression, a single nn.Linear layer with one input feature and one output feature suffices, instantiated as nn.Linear(1, 1) to reflect the model's input and output dimensions.

Forward Method: The forward method is where the actual computation of the model happens. It takes an input tensor x and passes it through the various layers of the model, in this case, the nn.Linear layer defined in the constructor. The forward method is automatically invoked when you call the model instance with an input, as in model(x).

Model Initialization: To create an instance of your linear regression model, you simply instantiate the class you've defined, which might be named LinearRegressionModel, by calling model = LinearRegressionModel(). This instance encapsulates your model's architecture and parameters, and it's ready to be trained on your data.

This setup provides a solid foundation for experimenting with linear regression in PyTorch. By leveraging the nn.Module class and its associated components, you can efficiently build, train, and evaluate linear regression models tailored to your specific data and tasks.

nn.Module: This is PyTorch's base class for all neural network modules. Your model should subclass this class.

init Function: This is the constructor for your class. Here, you initialize the layers of the model. In this case, there is only one layer - nn.Linear.

nn.Linear: This represents a linear transformation. For a simple linear regression, you need a single linear layer with one input feature and one output feature (hence nn.Linear(1, 1)).

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

Forward Method: This method defines the forward pass of the model. It takes an input tensor x and passes it through the linear layer.

Model Initialization: model = LinearRegressionModel() creates an instance of your linear regression model.

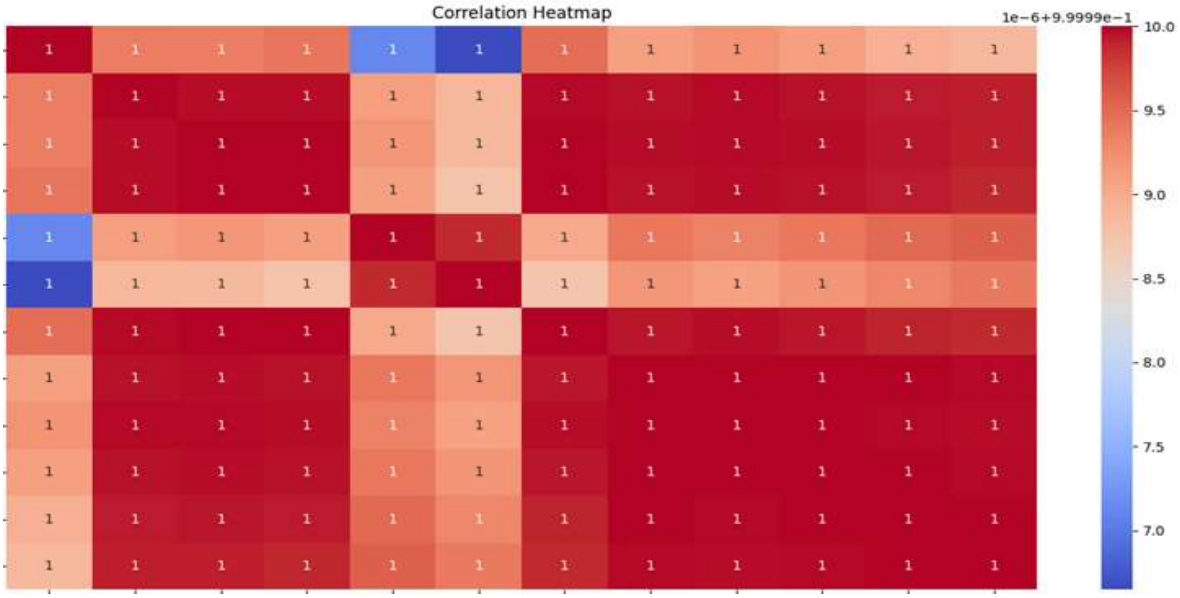

Figure 2: Co-relation Matrix

Each cell in the heatmap corresponds to the correlation coefficient between two variables. The correlation coefficient ranges from -1 to 1. A value of 1 implies a perfect positive correlation, meaning as one variable increases, the other variable also increases. A value of -1 implies a perfect negative correlation, meaning as one variable increases, the other variable decreases. A value of 0 implies no correlation between the variables.

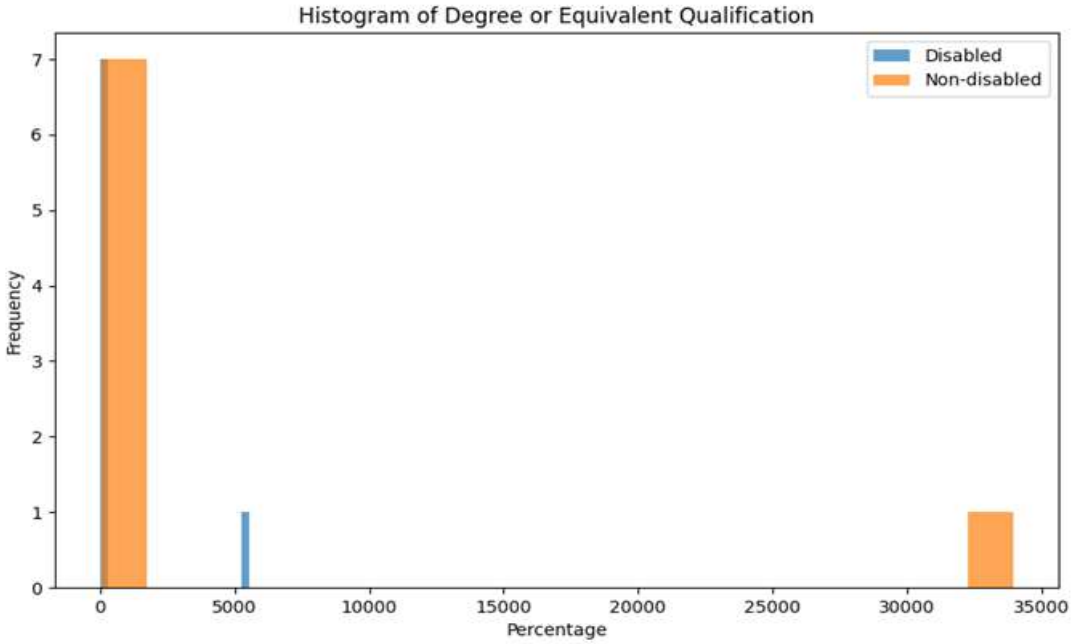

Figure 3: Histogram of degree or equivalent feature

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

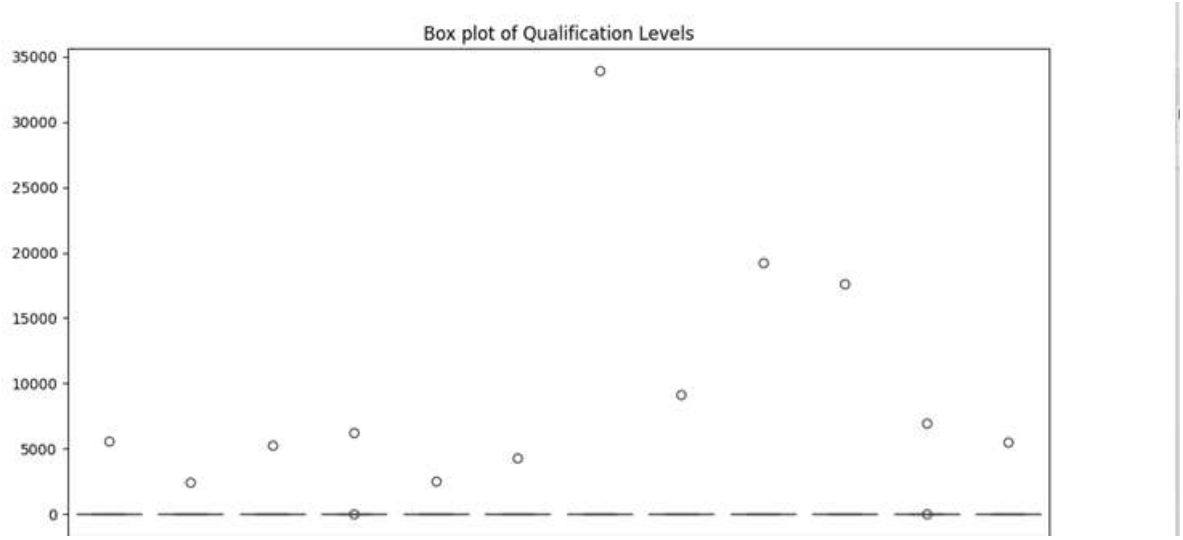

Figure 4: Box Plot of Qualification Levels

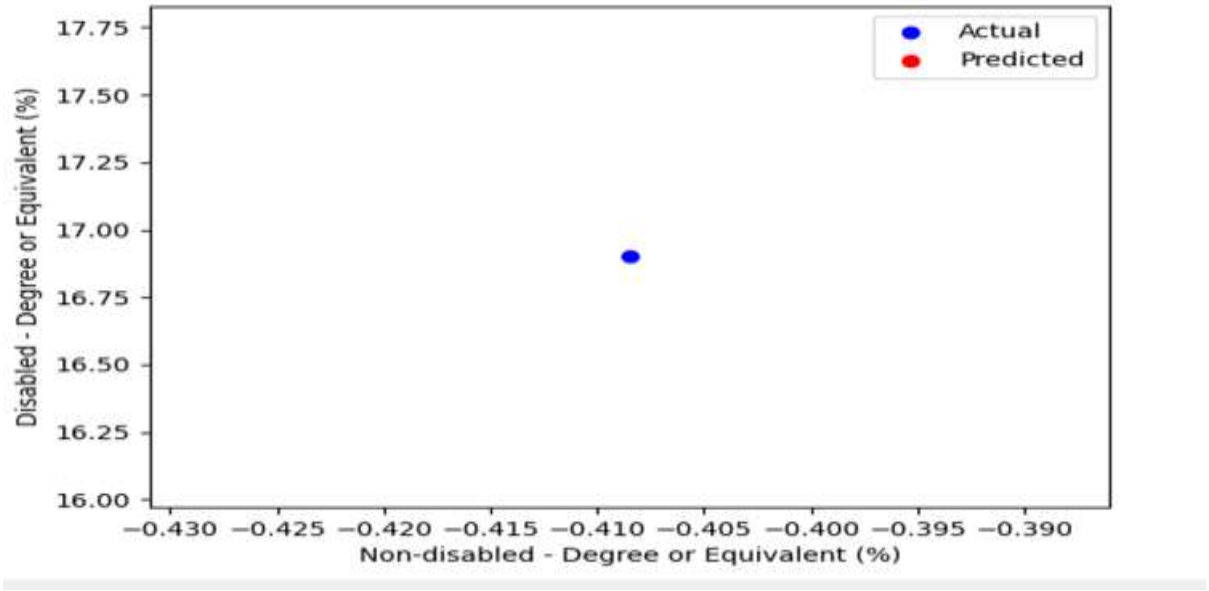

Figure 5: Actual and Predicted

B. Performance estimation in the first iteration

In the initial stages of model development and performance estimation, understanding the interrelationships between variables is paramount. This is particularly true in the context of linear regression and other statistical modelling techniques, where the strength and nature of these relationships can significantly influence model performance and interpretability.

Strong correlations (both positive and negative) are of interest because they indicate potential relationships between variables that may be worth further investigation.

Pattern Recognition: Patterns in the heatmap can reveal clusters of variables that are related to each other, which can be useful in feature selection for machine learning models, for identifying factors in multivariate analyses, and for reducing dimensionality in large datasets.

Diagonal Values: The diagonal values, which are all 1s, confirm that each variable is perfectly correlated with itself, as expected.

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

Potential Multicollinearity: If the heatmap were to show very high correlations between different independent variables, this could indicate multicollinearity, which can affect the performance of some statistical models by making the estimates of the coefficients unstable or hard to interpret.

Data Quality and Relevance: A correlation matrix can also help in assessing the quality and relevance of the data

This table 1 synthesizes key insights from previous research, highlighting the role of psychological frameworks, machine learning, and AI in understanding and improving educational outcomes. The implications of these studies underscore the potential for integrating advanced technologies and psychological insights to create more effective, personalized, and inclusive educational experiences. Future research directions and contributions are suggested to further explore these intersections, with an emphasis on leveraging technology to address gaps and challenges in educational psychology and pedagogy.

Table 1: Factors for educational outcomes

Aspect	*Focus*	*Key Models/Theories/Algorithms/Techniques*	*Methodology*	*Significant Authors/ Findings*	*Implications*	*Future Research/ Contribution*	*Identified Gaps/ Proposal*
1	Psychological well-being and SRL in university students	Keyes' model of mental health, Howell's and Hills & Argyle's work on emotional stability	Online tools for planning and reflection	Keyes, Howell, Hills & Argyle, Zimmerman, Winne & Hadwin, Schunk & Greene, Butler & Winne, Molenaar	Mental health and educational practices in higher education	Domains and predictors of student success, text-learning processes, motivation (Pintrich, Rogiers)	The Dataset quality and better classifications-
2	Predicting student academic performance using machine learning algorithms	Random Forest, SVM, Gradient Boosting, Decision Tree, Logistic Regression, XGBoost, Deep Learning	Analysis of historical academic data	XGBoost achieved high accuracy (97.12%)	Applications in educational settings for informed decision-making	Showcasing the potential of diverse machine learning algorithms in educational data mining	Research on further ensemble learning methods to validate the accuracy
3	Use of machine learning in predicting student performance	Classification methods, decision trees	Comprehensive datasets covering diverse levels and incorporation of psychological, environmental factors	We found this topic with suitable research potential	Improvement in educational data analysis and predictive modelling	We improved on this by creating a predictive performance system	Research on basic academic performance and intervention strategies
4	AI application in studying the impact of learning disabilities and study strategies on performance	Artificial Neural Networks (ANN), Fuzzy-based AI	Analysing the interplay between study strategies, learning disabilities, and academic performance	We found this topic with suitable research potential	Enhancing teaching methodologies, promoting inclusivity and equity in education	Predicting quantitative response variable based on one or more predictor variable can improve the	Application of AI in education,particularly for students with learning disabilities

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

						overall prediction algorithm	

Table 2 describes "Dataset" component and their description that indicates the presence of two distinct datasets, distinguished by different coloured bars. These bars likely represent different categories or groups within the data, such as non-disabled and disabled individuals. Moving on, the "X-Axis" denotes the variable being measured or categorized, which in this case pertains to specific degree categories or equivalent qualifications. There's a mention that this axis might be mislabelled as "Percentage," suggesting a possible error in labelling that needs correction. Next, the "Y-Axis (Frequency)" refers to the number of observations falling within each bin or category on the histogram. The "Bins" component signifies the division of the histogram into ranges of values, helping to organize the data for analysis. Finally, the "Distribution" aspect highlights how the data is distributed across these bins, providing insight into the concentration of observations within specific value ranges for each group represented in the dataset.

Table 2: Component Vs Description

Component	Description
Dataset	Two data sets represented by different coloured bars (e.g., blue for non-disabled, orange for disabled)
X-Axis	Represents a specific degree category or equivalent qualification, possibly mislabelled as "Percentage"
Y-Axis (Frequency)	Number of observations in each bin
Bins	Divide the histogram into ranges of values
Distribution	Indicates the concentration of data within specific ranges for each group

Here Table 3 is presenting the comparison of actual and predicted datasets for disabled and non-disabled students.

Table 3: comparison of actual and predicted datasets for disabled and non-disabled students

Type	Non-disabled - Degree or Equivalent (%)	Disabled - Degree or Equivalent (%)
Actual	0.625	0.571429
Predicted	0.625	0.514286

This table provides a concise explanation of the components involved in understanding a box plot visualization for qualification data.

5. Conclusion

The extensive research outlined in these references, including contributions from authors like Sekeroglu et al. [18] and Chair et al. [19], underscores the considerable advancements made in applying machine learning and data mining techniques to forecast and scrutinize student academic

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

performance. These studies collectively assess the effectiveness of various algorithms, the impact of cognitive, non-cognitive, and demographic factors, and the obstacles encountered in current methodologies. The findings, as discussed by authors such as Anuradha et al. [18] and Chair et al. [19], suggest that embedding self-regulated learning components into educational platforms, along with employing machine learning models to customize learning experiences, could significantly improve educational outcomes. The call for further investigation to refine the proposed model and its real-world educational applications is emphasized.

Expanding upon this groundwork, the exploration into advanced analytical techniques like Artificial Neural Networks (ANNs) for pattern identification, as highlighted by Sameer et al. and Li et al. [9], opens new pathways for deciphering complex data sets. The capacity for AI to customize interventions based on individual learning profiles, as proposed by Anuradha and Velmurugan [16] and Al-Radaideh et al. [23], foresees an educational future that is more adaptable and attuned to student necessities. Moreover, the integration of big data analytics, as explored by Garg et al. [24] and Garcia et al. [25], could reveal insights and tendencies obscured by conventional analyses, enabling a deeper comprehension of academic success determinants. Our study does comparative analysis of educational attainment between groups with and without disabilities based on region and diversity with around 90% accuracy. The findings indicate a strong link between self-regulated learning and academic success. The analysis explores the implications of these results for educational policies and practices, emphasizing the need for learning platforms to integrate features that facilitate self-regulated learning by Anuradha et al. and Chair et al. [18][19].

The future of machine learning and educational data analytics appears promising. The ongoing adoption of more sophisticated machine learning methods, like deep learning, promises to improve the precision of predictions and more adeptly manage intricate educational data sets. There is an increasing focus on utilizing extensive datasets that span various educational stages and incorporate both psychological and environmental elements, offering a fuller perspective of student achievement

As this domain progresses, the ethical considerations and privacy concerns surrounding educational data use, as cautioned by Lazzeroni and Lu [21] and Fernández-Delgado et al. [22], must be meticulously navigated. Safeguarding data security and upholding student confidentiality become paramount as these technologies gain traction within educational frameworks. Additionally, the creation of accessible platforms that both educators and students can easily utilize, as advocated by Romero and Ventura [27] and Okoli [28], is crucial for the broad acceptance of these innovations.

In conclusion, the path toward fully leveraging the potential of machine learning and AI in education continues, with significant strides already made by researchers like Barro et al. [31] and Jacob et al. [33]. Future research should extend beyond technological advancements to encompass the wider educational ecosystem, including policy ramifications, teacher training, and equitable technology access, fostering an environment where every student has the opportunity to excel. They highlight the growing importance of educational data mining in understanding and improving student outcomes, underscoring the potential of advanced analytics in the educational sector. The diversity of approaches and findings in these studies provides a comprehensive overview of this evolving field. Figure 1: This figure likely outlines the methodology of the study, including the steps taken during the research process. The methodology could involve data collection, analysis techniques, and the application of machine learning models to understand the impact of disabilities on educational attainment. Figure 2 (Correlation Matrix Heatmap): This figure shows the strength and direction of relationships between various variables studied. High positive values indicate strong positive correlations, meaning that as one variable increases, the other also tends to increase. Conversely, high negative values indicate strong negative correlations, where one variable tends to decrease as the other increases. Values close to zero suggest little to no linear relationship between the variables. Figure 4 (Histogram): The histogram depicts the distribution of a 'degree or equivalent' qualification among the study population, likely showing how many individuals fall into various categories of educational attainment. This can help in understanding the prevalence of certain levels of education within the groups studied. Figure 5 (Box Plot): This figure visualizes the distribution of qualification levels across different groups, possibly comparing those with and without disabilities. Box plots show the median, interquartile range, and potential outliers within the data, providing a clear picture of the central tendency and variability of qualification levels.

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

Dataset

Annual data on the highest level of qualification attained by disabled and non-disabled people aged 21 to 38 years. Office for National Statistics. (2019). Disability and education, UK. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/disability/bulletins/disabilityandeducationuk/2019

References

[1] Issah, I., Appiah, O., Appiahene, P., &Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. Data in Brief. https://doi.org/10.1016/j.dajour.2023.100204

[2] Balaji, P., Alelyani, S., Qahmash, A. (2021). Contributions of machine learning models towards student academic performance prediction: A systematic review. Applied Sciences.

[3] M. Nazir , A. Noraziah , M. Rahmah , Aditi Sharma. (2023). Examining the potential of machine learning for predicting academic achievement: A systematic review. Journal of Fusion: Practice and Applications, 13 (2), 71-90 (Doi : https://doi.org/10.54216/FPA.130207)

[4] Baashar, Y. (2021). Predicting Student's Performance using Machine Learning Methods: A Systematic Literature Review.

[5] Sandra, L., Lumbangaol, F., Matsuo, T. (2021). Machine Learning Algorithm to Predict Student's Performance: A Systematic Literature Review, 1919-1927.

[6] Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., Idoko, J.B. (2021). Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies. Applied Sciences.

[7] Anita Venugopal, Aditi Sharma, F. Abdul Munaim Al Rawas, Rama Devi S.. (2023). Enhancing Fusion Teaching based Research from the Student Perspective. Journal of Fusion: Practice and Applications, 12 (2), 109-119 (Doi : https://doi.org/10.54216/FPA.120209)

[8] Nawang, H., Makhtar, M., Hamzah, W.M.A.F.W. (2021). A systematic literature review on student performance predictions. International Journal of Advanced Technology and Engineering Exploration, 8(84), 1441-1453.

[9] S. Phani Praveen, Kotte Sandeep, N. Raghavendra Sai, Aditi Sharma, Jitendra Pandey, Vikas Chouhan. (2024). Outlier Management and its Impact on Diabetes Prediction: A Voting Ensemble Study. Journal of Journal of Intelligent Systems and Internet of Things, 12 (1), 08-19 (Doi : https://doi.org/10.54216/JISIoT.120101)

[10] Albreiki, B., Zaki, N., Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. Educational Sciences, 11(9).

[11] Davis, S. K., & Hadwin, A. F. (2021). Exploring differences in psychological well-being and self-regulated learning in university student success. Frontiers in Language Research, 9(1). https://doi.org/10.14786/flr.v9i1.581

[12] Rahul Sharma , Shiv Shakti Shrivastava , Aditi Sharma. (2023). Predicting Student Performance Using Educational Data Mining and Learning Analytics Technique. Journal of Journal of Intelligent Systems and Internet of Things, 10 (2), 24-37 (Doi : https://doi.org/10.54216/JISIoT.100203)

[13] NoorUlhuda S. Ahmed, Esraa H. Alwan, Ahmed B. M. Fanfakh. (2024). Optimizing Loop Tiling in Computing Systems through Ensemble Machine Learning Techniques. Journal of Fusion: Practice and Applications, 15 (1), 214-226 (Doi : https://doi.org/10.54216/FPA.150117)

[14] Li, L., Yang, D., & Xu, K. (2020). A deep learning-based method for predicting academic performance of students. Journal of Ambient Intelligence and Humanized Computing, 11(2), 591-598.

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

[15] Ofori, F., Maina, E., Gitonga, R. (2020). Using machine learning algorithms to predict students' performance and improve learning outcome: A literature-based review. Journal of Information Technology, 2617-3573, 4(1), 33-55.

[16] Enughwure, A.A., Ogbise, M.E. (2020). Application of Machine Learning Methods to Predict Student Performance: A Systematic Literature Review, 3405-3415.

[17] Gao, F., Luo, T., & Zhang, K. (2018). A predictive analytics approach for academic performance prediction of college students. IEEE Access, 6, 73443-73454.

[18] Bista, S. K., & Gauch, S. (2017). Improving academic performance prediction using data mining techniques: A case study of an engineering college in Nepal. International Journal of Information Technology and Electrical Engineering, 6(3), 9-16.

[19] Makhtar, M., Nawang, H. (2017). Analysis on students' performance using naïve Bayes classifier. Analysis on Students Performance using NAÏvE.

[20] Pandey, M., Taruna, S. (2016). Towards the integration of multiple classifiers pertaining to the student's performance prediction. Perspectives in Science, 8, 364-366.

[21] Anuradha, C., Velmurugan, T. (2016). Fast Boost Decision Tree Algorithm: A Novel Classifier for the Assessment of Student Performance in Educational Data, 31, 254–0223.

[22] Sumitha, R., Vinothkumar, E.S., Scholar, P.G. (2016). Prediction of student's outcome using data mining techniques. International Journal of Science Engineering and Applied Sciences, 2(6), 132-139.

[23] Dilobar Isomjonovna Ruzieva. (2024). The Fusion of Digital Technologies in Small Business for Ensuring the Socio-Economic Development: Panel Data Analysis. Journal of Fusion: Practice and Applications, 15 (1), 66-77 (Doi : https://doi.org/10.54216/FPA.150106)

[24] Anuradha, C., Velmurugan, T. (2016). Feature selection techniques to analyse student academic performance using naïve bayes classifier. 3rd International Conference on Small Medium Business, 345-350.

[25] Chair, P., et al. (2016). Organizing Committee General and Financial Chair Organizing Secretary Technical Programme Committee Advisory Committee Predicting and Analyzing Students' Performance: An Educational Data Mining Approach.

[26] Sameer, P.G., Barahate, S.R. (2016). Educational Data Mining – A New Approach to the Education Systems, 18-20.

[27] Lazzeroni, L., & Lu, Y. (2016). Belief propagation: an iterative method for variable selection in high-dimensional linear regression. Journal of Machine Learning Research, 17(1), 263-279.

[28] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research, 15(1), 3133-3181.

[29] Al-Radaideh, Q. A., & Al-Shawakfa, E. M. (2014). A hybrid intelligent method for predicting students' performance using machine learning algorithms. International Journal of Database Theory and Application, 7(4), 89-102.

[30] Garg, D., & Rani, R. (2014). An efficient classification technique for predicting student's performance. International Journal of Computer Applications, 86(7), 29-33.

[31] García, V., Mollineda, R. A., Sánchez, J. S., & Alejo, R. (2012). Predicting student failure at school using genetic programming and different data mining approaches. Neurocomputing, 89, 128-138.

[32] Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). Learning from data. New York, NY: AMLBook.

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024

[33] Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.

[34] Okoli, C. (2015). A Guide to Conducting a Standalone Systematic, 37.

[35] Pandey, M., Taruna, S. (2016). Towards the integration of multiple classifiers pertaining to the student's performance prediction. Perspectives in Science, 8, 364-366.

[36] Winne, P. H. (2006). How software technologies can improve research on learning and bolster school reform. Educational Psychologist, 41(1), 13.

[37] Barro, Robert J., & Lee, Jong-Wha. (2010). "A New Data Set of Educational Attainment in the World, 1950-2010." This paper presents an updated dataset on educational attainment across 146 countries, offering a comprehensive look at global education trends which could be instrumental in comparative studies on disability and education. Available at NBER: www.nber.org.

[38] Reber, Sarah J. (2010). "School Desegregation and Educational Attainment for Blacks." This study explores the impact of school desegregation on black students in the U.S., focusing on changes in educational funding and exposure to diverse student populations. The findings could provide a historical perspective on how systemic changes in education impact minority groups, which might parallel some challenges faced by individuals with disabilities. Available at NBER: www.nber.org.

[39] Jacob, Brian A., & Wilder, Tamara. (2010). "Educational Expectations and Attainment.": www.nber.org.

Doi: https://doi.org/10.54216/FPA.150204
Received: August 17, 2023 Revised: December 06, 2023 Accepted: March 21, 2024