JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Classifying School Scope Using Deep Neural Networks Based on Students' Surrounding Living Environments

Mochammad Daffa Putra Karyudi and Anis Zubair *

Faculty of Technology Information, Universitas Merdeka Malang, Indonesia;
e-mail : m.daffa.karyudi@gmail.com; anis.zubair@unmer.ac.id
* Corresponding Author : Anis Zubair

Abstract: This research investigates school scope classification using Deep Neural Networks (DNN), focusing on students living environments and educational opportunities. By addressing the interplay of socioeconomic and educational factors, the study aims to develop an analytical framework for understanding how environmental contexts shape academic trajectories. The research provides a nuanced understanding of the importance of features in educational classification by developing DNN models based on Spearman's Rank Correlation Coefficient (SRCC). The methodology employs machine learning techniques, integrating data wrangling, exploratory analysis, and multiple DNN models with K-fold cross-validation. The study analyzes 677 student records from two schools. The research examined multiple model configurations. Results show that the 'All Data' model achieved 83.08% accuracy, the 'Top 5' model 81.54%, and the 'Non-Top 5' model 79.23%. The SRCC-based approach revealed that while top correlated features are important, additional variables significantly contribute to model performance. The study highlights the profound impact of family background, social environment, and educational contexts on school selection. Furthermore, it demonstrates DNN's capability to uncover intricate, non-linear relationships, offering actionable insights for policymakers to leverage machine learning's potential in developing targeted educational strategies.

Keywords: Classification; Deep Neural Networks; Educational Analytics; Educational Interventions; Overfitting Prevention; School Scope.

1. Introduction

The environments in which students live profoundly shape their educational experiences. Family and friends define the closest aspects of students' lived environments, mirroring research on how a student's surroundings can impact their learning process [1]–[3]. These studies demonstrate that the environment surrounding students can influence their educational outcomes. In line with this, classification techniques allow researchers to understand better how factors such as learning environment and social interaction affect educational success[4]. Given the importance of school in students' lives[5], attention to these factors is essential.

Understanding these complex dynamics has become increasingly relevant, especially with the importance of family, friends, and school environments in influencing students' educational process. Researchers have harnessed the deep data analysis capabilities of advanced technologies like Machine Learning (ML) to automate the identification of factors that influence students' educational development[6]. This has enabled a more structured understanding of the intricate environmental factors. Therefore, effective learning methods that match the characteristics of the data being analyzed are required to optimize the utilization of such approaches. While traditional ML models like Support Vector Machines(SVM) and Random Forests (RF) provide linear interpretations, they have limitations in capturing the complex. These non-linear relationships influence students' educational trajectories[7].

In the evolving landscape of educational research, applying advanced machine learning technologies has become increasingly relevant. Particularly, Deep Neural Networks (DNN) have emerged as a powerful approach to understanding complex educational dynamics[8].

Received: November, 2nd 2024
Revised: November, 25th 2024
Accepted: November, 27th 2024
Published: November, 28th 2024

Unlike traditional ML methods or Recurrent Neural Network(RNN) models like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), DNNs offer unique capabilities in processing multidimensional data related to students' educational environments[9]. The choice of DNNs is fundamentally rooted in their ability to model intricate, non-linear relationships that conventional algorithms struggle to capture. This capability allows DNNs to identify much more complicated patterns than traditional machine learning algorithms, making this technology ideal for applications requiring high reliability and precision[10]. Supervised Learning methods, for example, involve training a model with a pre-labeled dataset to learn the relationship between known inputs and outputs, facilitating its use in various sectors[11].

Spearman's Rank Correlation Coefficient (SRCC) is often used to analyze the relationship between variables in research contexts that do not rely solely on normal data distribution. The SRCC is effective for measuring the strength and direction of a monotonic relationship between two ordered variables, making it a very practical tool, especially in studies involving ranked data[12]. However, using SRCC also requires a normality test to ensure its suitability, whereas the Shapiro-Wilk test is recommended due to its ability to detect deviations from normality[13].

This study makes significant contributions to educational analytics. First, it introduces a new way to use advanced ML, specifically DNN, to classify school scope based on students' neighborhood environment, an area not widely studied before. Second, the research focuses on environmental factors like family background and social environment, providing a deeper, more complete understanding of how these variables influence school selection filling an important gap in existing literature. Third, the study integrates advanced feature selection strategies and uses a multi-model approach, systematically employing SRCC to examine the impact of different variable combinations. This comprehensive method uncovers intricate patterns and complex, non-linear relationships that traditional approaches may miss. Ultimately, the research offers a sophisticated analytical framework that can inform policymakers and educators, potentially leading to targeted interventions and improved educational outcomes by considering the broader social context in which students live.

2. Related Work

The exploration of statistical tests and machine learning techniques has been pivotal in various research domains, including educational studies and reliability analysis. The simulation study compared the Kolmogorov-Smirnov, Skewness, and Shapiro-Wilk normality tests, emphasizing that selecting an appropriate normality test depends on the sample size and underlying data distribution[14]. This insight is particularly relevant to our study, where the Shapiro-Wilk test assesses the normality of variables influencing school choice.

Research highlighted the critical role of reliability in the design of DNN algorithms and their accelerators[8]. Mittal underscored that reliability should be a fundamental consideration across all abstraction levels in system design when utilizing classification and DNN methodologies. This perspective aligns with our approach, which integrates DNN alongside other machine learning techniques to ensure robust classification of school scopes based on students' living environments.

A review of various normality tests, including Shapiro-Wilk, Shapiro-Francia, Anderson-Darling, Cramer-Von Mises, Lilliefors, and Jarque-Bera, concluded that the Shapiro-Wilk, Shapiro-Francia, and Anderson-Darling tests exhibit superior power compared to others[13]. This reinforces our methodological choice to utilize the Shapiro-Wilk test in analyzing the significance of life variables on school selection.

Implementation of machine learning techniques, specifically supervised learning to predict student performance, achieved an accuracy rate of over 80%, demonstrating the efficacy of machine learning in predictive tasks[6]. These precedent supports using supervised learning within our machine learning framework to classify school scopes accurately.

Investigating the correlation between personality traits and visiting places using SRCC, Pearson's linear correlation coefficient (PLCC), and Kendall's rank correlation coefficient (KRCC) indicated that SRCC provided more reliable outcomes than the other methods[12]. This finding justifies our application of SRCC to explore the relationship between students' living environments and school choices.

Employed various machine learning algorithms, including Naive Bayes (NB), Logistic Regression (LR), and SVM, alongside K-fold cross-validation, to effectively classify opinions regarding university student satisfaction[15]. Their research demonstrated that SVM, combined with K-fold cross-validation, successfully identified key factors influencing satisfaction. This methodology is mirrored in our study, where we utilize K-fold cross-validation alongside SVM to enhance the classification accuracy of school scopes based on environmental variables.

Unlike previous studies focused on predicting student performance using traditional machine learning, this research takes a novel approach by using a comprehensive DNN to classify school scope based on students' surrounding living environments. While previous research has used data mining techniques like Decision Trees (DT), RF, and SVM to predict student achievement[16], our study investigates how environmental factors influence school selection.

Table 1. Previous Research.

Ref.	Method	Result
[14]	Kolmogorov-Smirnov, Skewness and Shapiro-Wilk.	Shapiro-Wilk test gives better results than other tests for normal and non-normal distributions, and increasing sample sizes increases test power.
[8]	Classification using DNN	Reliability must be considered a key design consideration at all levels of abstraction.
[13]	Shapiro Wilk, Shapiro Francia, Andeson Darling, Cramer Von Mises, Lilliefors, and Jarque Bera.	The Shapiro Wilk test, Shapiro Francia test, and Andeson Darling test are the most powerful among the other tests.
[6]	ML, Supervised Learning, DT	One Rule, Joint Reserve Intelligence Program, and DT accurately predict student academic performance with over 80% accuracy.
[12]	SRCC, PLCC, and KRCC.	SRCC showed better results than PLCC and KRCC.
[15]	ML: NB, LR, SVM, and K-fold cross-validation.	SVM classification using K-fold cross-validation will contribute to determining which factors related to teachers' didactic strategies should be improved.
[16]	Classification using DT, RF, Neural Network, SVM	The study demonstrates the effectiveness of data mining techniques in predicting secondary student performance, emphasizing the importance of previous academic results and various demographic and social factors.

The exploration of educational analytics has witnessed significant advancements in recent years, with researchers employing various ML techniques to understand complex educational dynamics. The studies in Table 1 provide a solid basis to our research, highlighting the critical need to choose suitable statistical tests and machine learning approaches to analyze and predict complex educational relationships.

Building upon methodological innovations in educational data mining, our research integrates advanced feature selection strategies. For instance, the Local Lift Dependence Scale (LLDS) demonstrated the potential of multi-resolution feature selection, extending beyond classical approaches by analyzing local dependencies between variables[17]. Similarly, utilizing K-mean clustering revealed nuanced patterns in student behavior, highlighting the complexity of educational performance prediction[18].

The proposed methodology extends beyond the conventional performance prediction models by leveraging advanced supervised learning techniques. Where studies that explored probabilistic evaluation of classification models[19], our research introduces a multi-model approach that systematically examines the impact of different variable combinations. Developing DNN models based on the result of SRCC provides a more nuanced understanding of the importance of features in educational classification.

Complementing neural network approaches, researchers applied Back Propagation and Radial Basis Function Networks to predict student performance with recognition rates exceeding 84%[20]. Our study innovatively addresses limitations by demonstrating that school scope determination is not merely a function of academic performance but a complex interplay of socioeconomic, familial, and environmental factors.

Unlike the wrapper techniques for handling imbalanced datasets proposed by previous researchers[21], and the penalized regression methods explored for improving research replicability[22], we integrate a comprehensive feature selection strategy beyond traditional reweighting and regularization methods.

The study's unique contribution lies in its ability to process and analyze multidimensional data using DNN, revealing intricate patterns that conventional ML algorithms struggle to capture. Focusing on the relationship between students' living environments and school scope offers a more holistic perspective on educational dynamics, potentially providing policymakers and educators with actionable insights for targeted interventions.

Critically, our approach differs from existing studies by predicting outcomes and understanding the underlying mechanisms that influence educational opportunities. The developed DNN model demonstrates superior capabilities in identifying complex, non-linear relationships between environmental variables and school selection, thus offering a more sophisticated analytical framework than traditional linear regression or classification techniques.

3. Proposed Method

3.1 Data Wrangling

The research begins with data wrangling, which involves converting raw data into a more analyzable format. This procedure encompasses data cleaning, integration, and transformation. The dataset used in this research encompasses a variety of attributes related to students' demographics, family background, and academic performance. For this data wrangling task, the powerful and flexible pandas package in Python was utilized[23]. Additionally, the scikit-learn library was employed for normalization[24].

Normalization is a crucial step, particularly for algorithms like neural networks, which are sensitive to the scale of input data. The scikit-learn library[24]. Provided various normalization techniques, including MinMaxScaler, StandardScaler, and OneHotEncoder, which were utilized in this study.

3.2 Exploratory Data Analysis

Exploratory Data Analysis is conducted to comprehend the dataset's fundamental structure and underlying patterns thoroughly. This comprehensive process involves computing descriptive statistics, such as the mean, median, mode, and other relevant metrics, to understand the data better. Additionally, the Shapiro-Wilk test is employed, leveraging the powerful scipy library in Python[25] to evaluate the normality of the data. This normality assessment helps determine the appropriate statistical techniques for the subsequent analysis.

3.2.1 Shapiro Wilk

The Shapiro-Wilk test is a normality test commonly used in frequency statistics. The null hypothesis of this test is that the population is normally distributed. If the p-value is less than the chosen alpha(a) level, the null hypothesis is rejected, indicating that the data being tested does not come from a normal distribution, i.e., the data is not normal. Conversely, if the p-value is greater than the selected level, the null hypothesis stating that the data comes from a normally distributed population cannot be rejected. This means the data follows a normal distribution and can be analyzed using statistical methods that assume normality, such as parametric tests[13].

The Shapiro-Wilk equation can be written using Equation (1)[14].

$$w = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1}$$

Where $a_i x_{(i)}$ represents the i-th data point in the sorted sample, where a_i is the sample size coefficient from the standard normal distribution table, and $x_{(i)}$ is the i-th smallest value in the sample. The term x_i refers to the i-th value of the sample data being tested, and $\bar{x}$ is the sample mean.

3.2.2 Spearman's Rank Correlation Coefficient (SRCC)

SRCC is a nonparametric statistic that measures the strength and direction of a monotonic relationship between two variables. Unlike Pearson's correlation, SRCC does not require a linear relationship between the variables. It is based on the ranks of the observations rather than the actual values.

To calculate SRCC, the observations of each variable are first ranked, and then the correlation between the ranks is computed. This makes SRCC a useful tool for analyzing relationships in data that do not follow a normal distribution, as it is based on the ranks rather than the actual data values. The calculation of this coefficient can be carried out using the Equation (2)[12].

$$p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{2}$$

Where x_i and y_i are the i-th values of the x and y sample data being tested, respectively, $\bar{x}$ and $\bar{y}$ are the means of all values in the x and y sample data, respectively.

3.3 Data Visualization

Data visualization techniques transform raw data into graphical representations that facilitate interpretation. Methods like heatmaps, pie charts, treemaps, and box plots accentuate key patterns and trends within the data. These visual tools offer a more intuitive comprehension of the underlying data, unveiling insights that may not be readily discernible from raw numerical information. For this purpose, the Matplotlib library in Python was utilized[26], leveraging its flexibility and extensive plotting capabilities to create highly customizable and informative visualizations.

3.4 Deep Neural Network

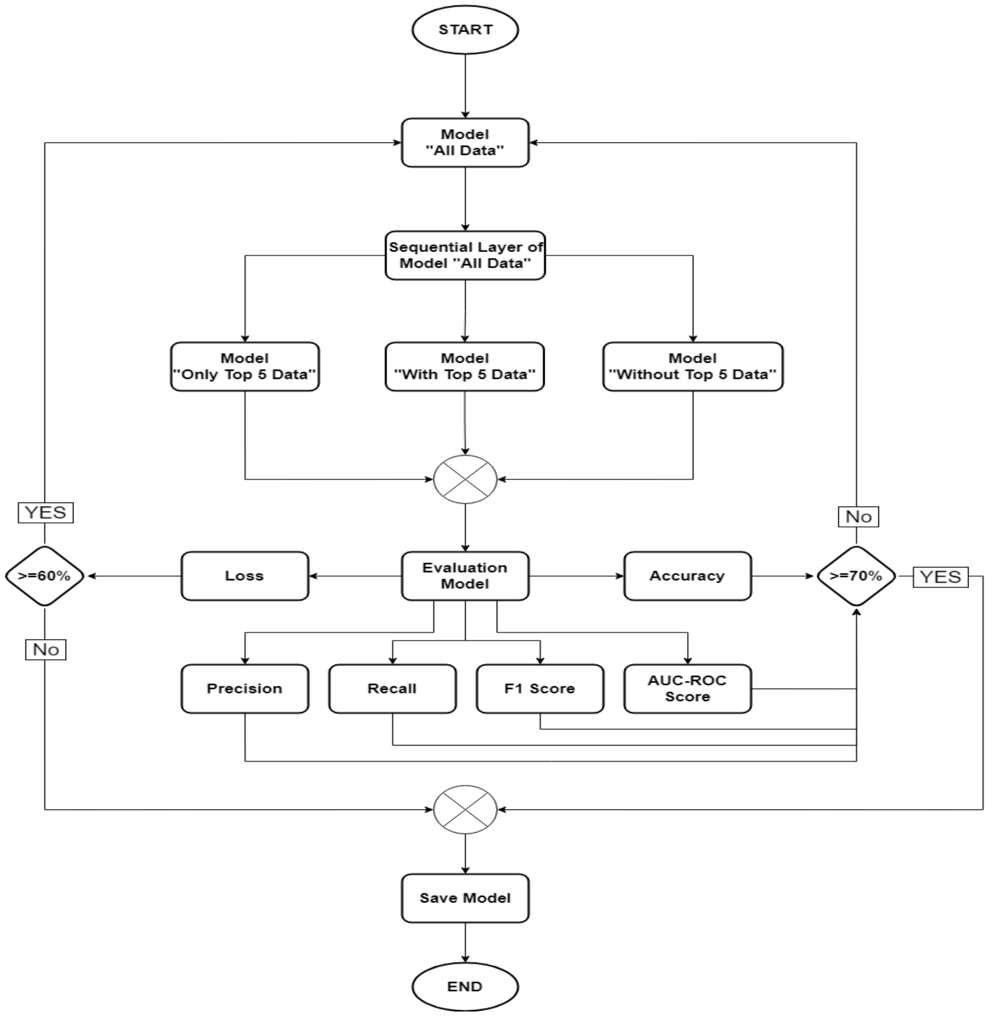

Figure 1. Deep Neural Network Model Plan

In this project, based on Figure 1. Four different DNN models to understand the influence of certain features on model performance and identify which model is most effective in various contexts. The data training process divided the dataset into two main segments: train data and test data, with a ratio of 80% and 20% for test data. This division aims to ensure that the model can learn from various possible scenarios in the dataset and be tested with independent data to verify the accuracy of the model's predictions.

In addition, to improve the model's reliability and reduce the risk of overfitting, a cross-validation technique, K-fold cross-validation, is applied during the training process. In K-fold cross-validation, the training data is divided into 'K' different subsets, with the model trained using 'K-1' subset as training data and the remaining subset as validation data. This technique allows utilization of the entire training data for training and validation, minimizing bias and providing a more stable estimate of the model's performance.

Main Model ("All Data" Model): This is the most comprehensive model, using the entire dataset without restrictions on the features used. This model aims to understand the entire dataset's performance in machine learning and train it to handle various scenarios in the dataset.

"Only Top 5 Data" model: This model uses exclusively the top five features identified as the most significant. It aims to test how well the model can perform with only the most crucial information.

"With Top 5 Data" model: Unlike the previous model, this model uses the entire dataset but specializes on the top five features. This model provides insight into how these important features interact with other features in the dataset.

The "Without Top 5 Data" model tests the dataset by removing the top five features to assess how effective the model is at predicting or classifying without relying on the most critical features.

After the training process, different domains prioritize distinct performance metrics. School scope classification demands a holistic multi-metric evaluation approach. This nuanced methodology reflects the complex nature of categorizing school environments, where understanding the contextual characteristics requires a comprehensive and multidimensional assessment.

Accuracy is the primary metric, capturing the model's overall correctness and providing a broad understanding of predictive capabilities. It offers stakeholders an intuitive and immediate insight into the model's reliability, reflecting the fundamental ability to make precise school scope classifications. By measuring the proportion of correct predictions across all instances, accuracy provides a foundational view of the model's effectiveness in understanding and categorizing school environment characteristics[19].

The loss metric provides insights into the model's internal performance, measuring prediction error and model uncertainty. It helps researchers understand the learning process, identifying potential overfitting or underfitting scenarios. By guiding model refinement and optimization, the loss metric serves as a critical tool for the continuous improvement of classification models[19].

Precision focuses on the accuracy of positive classifications, which is crucial in identifying specific school scope characteristics while minimizing false positive classifications. This metric ensures targeted and efficient categorization strategies, reducing potential misclassification risks. By carefully filtering predictions, precision helps researchers design more accurate environmental classifications[19].

Recall complements precision by capturing the model's ability to identify all relevant school scope instances. This metric ensures comprehensive identification of environmental characteristics and prevents overlooking critical contextual attributes. Recall supports a more comprehensive classification approach by minimizing missed detection of specific school environment categories [19].

The F1 score emerges as a critical metric that balances precision and recall, addressing the challenges inherent in imbalanced school scope classification datasets. By providing a harmonized view of model performance, the F1 score mitigates the limitations of single-metric evaluation. This is particularly valuable in contextual classification, where class distributions may be uneven, ensuring a comprehensive and balanced assessment of predictive capabilities[19].

The AUC-ROC score introduces a more sophisticated layer of analysis, demonstrating the model's ability to distinguish between different school scope categories. This metric reveals nuanced classification capabilities by measuring the model's discriminative power across various environmental thresholds. Unlike binary classification approaches, the AUC-ROC score helps researchers understand how well the model can rank and differentiate predictions, providing insights into the subtle variations in school environment characteristics[19].

This method aims to develop robust and reliable DNN models, capable of making accurate predictions under various conditions critical for real-world applications using machine learning technologies. Each model is evaluated based on performance metrics to determine which is most effective under different conditions. This allows optimization based on the most significant features for accurate predictions and effective classification.

4. Results and Discussion

This chapter presents the findings from the data analysis. The objective is to investigate the relationships and distributions of the diverse factors that influence the categorization of school scope based on students' surrounding living environments. By examining these factors, this study offers more profound insights into how the educational milieu can impact students' academic performance.

The research utilizes a UC Irvine ML Repository dataset, specifically the "Student Performance" dataset[27]. This dataset represents a meticulously curated collection of educational performance data, collected through a rigorous methodology involving school reports and a carefully designed supplementary questionnaire, undergoing a process of professional review and student piloting.

The original survey encompassed 788 students, with 111 responses ultimately excluded due to incomplete identification details, resulting in a refined dataset of approximately 677 student records. The questionnaire, strategically constructed with 37 closed-ended questions, comprehensively captured demographic, social/emotional, and school-related variables hypothesized to influence student performance.

Preliminary analysis suggests the dataset comprises a rich mix of approximately 15-20 categorical variables and 17-22 numerical variables, potentially including features such as gender, school type, parental education, grades, age, and study time. For a complete feature list and dataset characteristics, refer to [27].

The variable 'school' was used as the target for classification in this study. This variable includes two categories: 'GP' for Gabriel Pereira school and 'MS' for Mousinho da Silveira school. These two categories represent the educational institutions analyzed to evaluate and compare the environmental and demographic factors affecting students' performance.

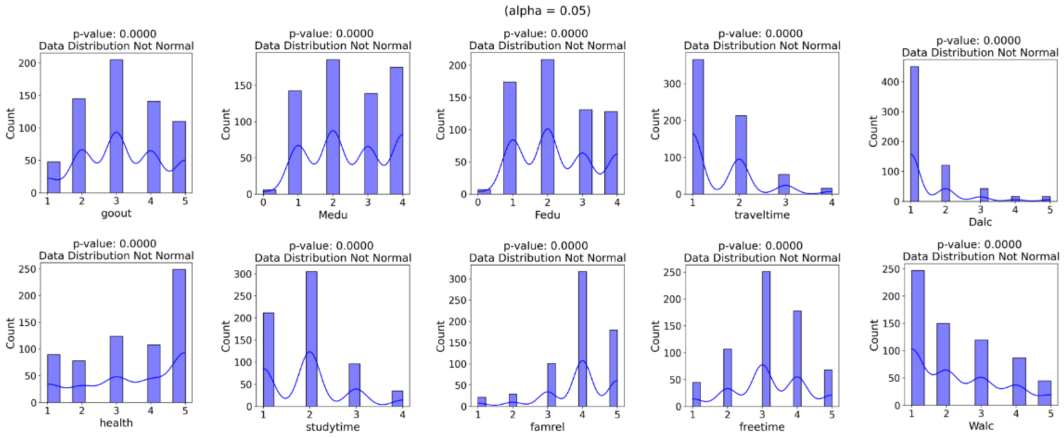

Figure 2. Shapiro Wilk of ranking data

The analysis of Figure 2 indicates that the data distribution does not conform to the normal distribution assumption, as evidenced by the p-value being lower than the set significance level. This suggests that the data does not meet the normality assumptions [14]. Disregarding the findings of the normality test and presuming normal distribution would be inadvisable, as it could result in inaccurate calculations and statistical inferences. When the data fails to meet the normality assumption, researchers should consider employing alternative non-parametric methods or data transformation techniques to ensure the validity and reliability of the analysis[13].

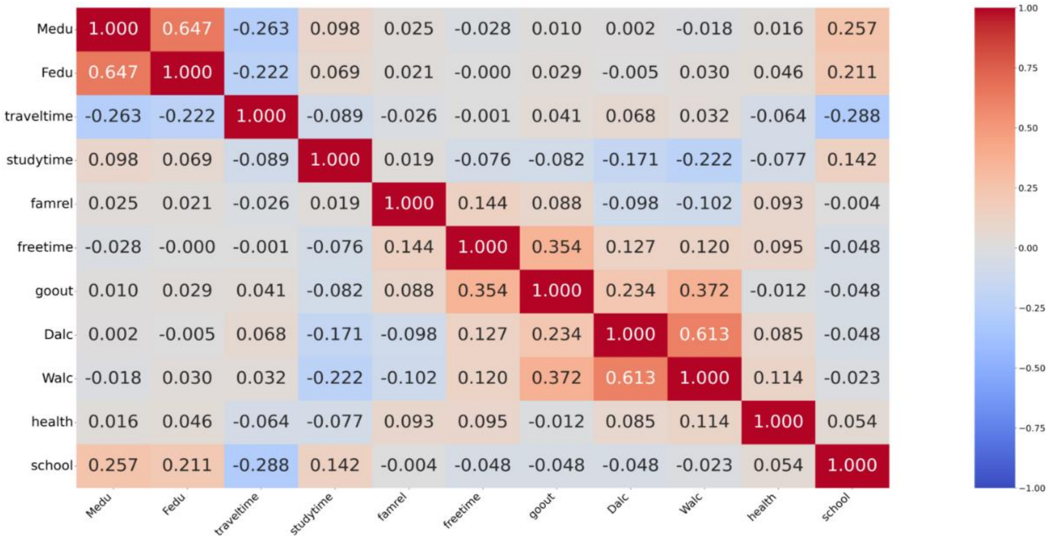

Figure 3. Heatmaps of correlations to 'school' data

Correlation heatmaps in Figure 3 offer a useful way to visualize the relationships between variables in the dataset. These heatmaps use color gradients to represent the strength and direction of correlations, where darker/lighter colors indicate stronger/weaker correlations. The scale from -1 to +1 denotes the range from a perfect negative to a perfect positive relationship, while values near 0 suggest no correlation. Heatmaps are valuable for identifying variables with significant influence on each other, enabling further focused analysis on those variables. This correlation information is crucial for evaluating the classification accuracy[28].

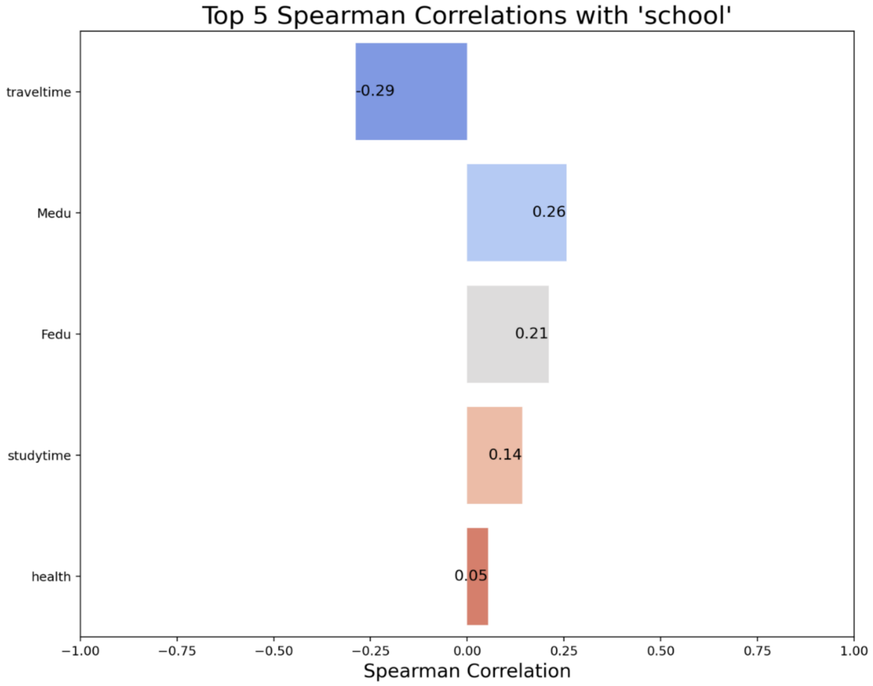

Figure 4. Top 5 Correlations with Target

Figure 4 shows the five variables most strongly linked to the target variable. Each bar indicates the strength of the correlation in order from highest to lowest. This identifies the key factors influencing the target, which is important for making informed decisions and developing strategies. Knowing the strongly correlated variables allows further study of their relationships, as seen in previous classification model research[19]. These insights can guide how to improve model training best.

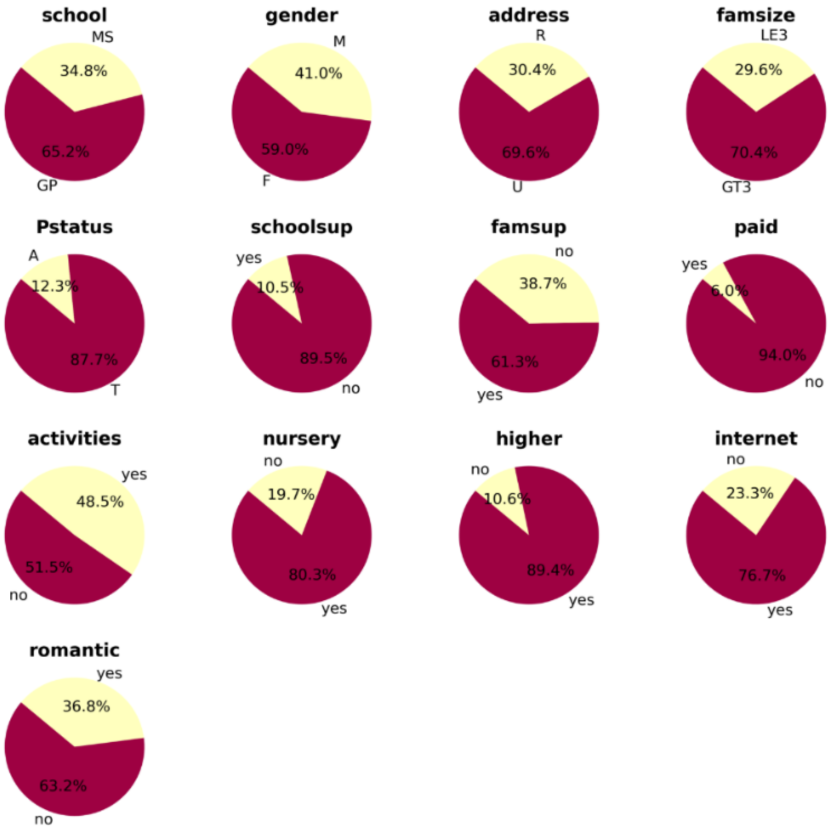

Figure 5. Pie Chart of Binary data

Figure 5 presents the proportional breakdown of various student attributes associated with their school environment. For instance, it indicates that most students attend the Gabriel Pereira school compared to the Mousinho da Silveira school, and there is a predominance of female over male students. This visual representation of the demographic composition facilitates comparisons across different student categories. As preliminary research [29] suggested, this type of visualization can provide valuable insights into the demographic factors that may influence educational policies and teaching approaches.

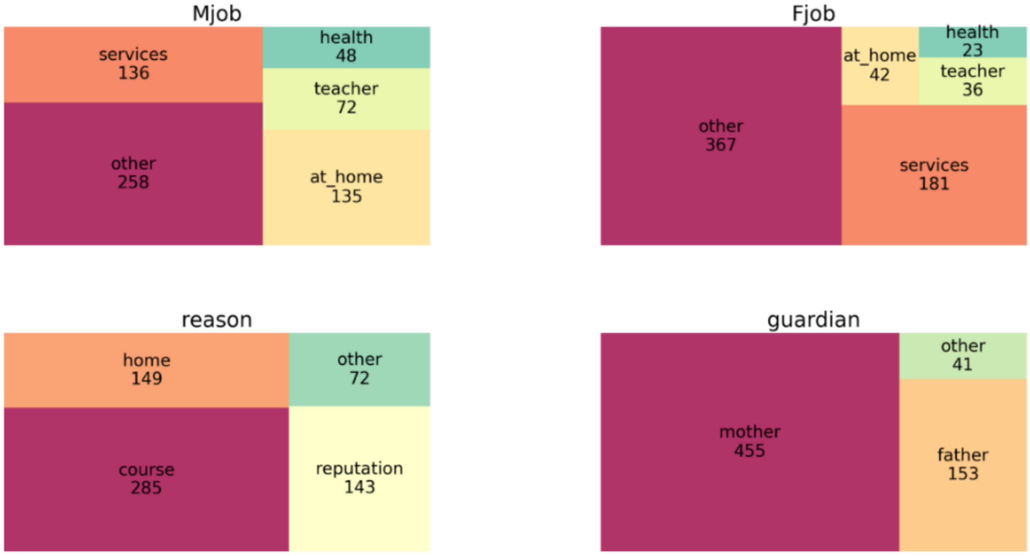

Figure 6. Treemap of Mjob, Fjob, reason, and guardian data

Figure 6 presents a visual breakdown of parents' occupational profiles and the primary reasons students select their schools. The color-coded blocks denote the relative proportions for each category. For instance, the "other" occupation category is the most prevalent for mothers and fathers. Additionally, most students cite the available course offerings as the primary factor in their school choice. This graphical representation provides valuable insights into the socioeconomic and academic factors that may shape students' educational decisions, aligning with research emphasizing the importance of understanding the social context in the educational domain[29].

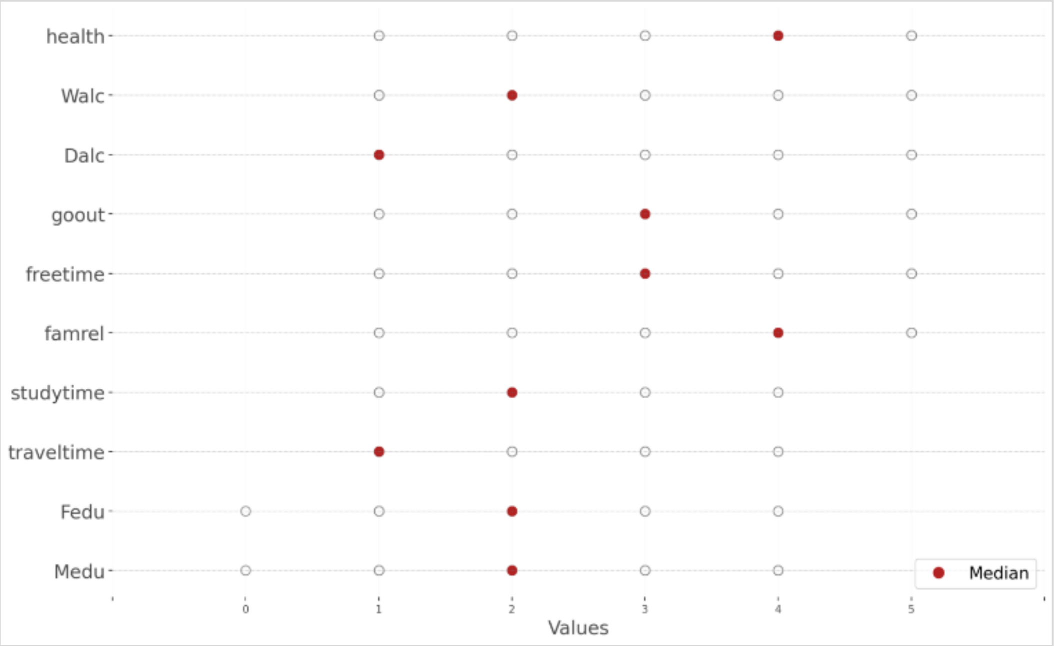

Figure 7. Distributed Data Plot of ranking data

Figure 7 presents the median values of variables related to students' health, alcohol consumption, and parental education levels, offering a snapshot of the general status of these attributes within the student population. This visualization serves to identify the overall condition of factors influencing the student's lives and highlight potential areas requiring targeted interventions. This technique facilitates the rapid detection of common patterns or prevalent issues in educational datasets[29], enabling more informed decision-making and the development of tailored strategies to address the student population's needs.

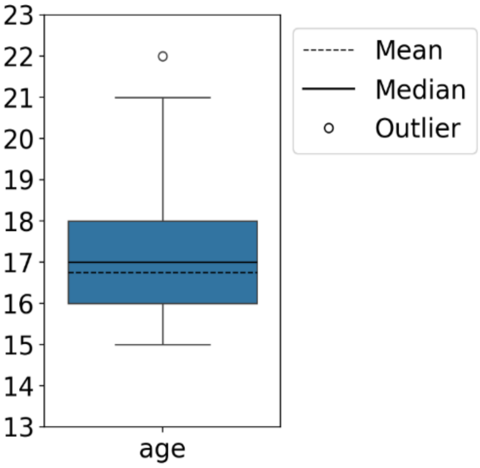

Figure 8. Boxplot from the age feature

The boxplot in Figure 8 depicts the age distribution of the student population, showcasing key statistical measures such as quartiles, medians, and outliers. The median student age is approximately 17, with some individuals exhibiting significantly higher ages. This visual representation effectively highlights the age distribution and exceptional cases that may warrant further investigation. As described in the literature[29], the utilization of boxplots aids in illustrating the statistical characteristics of the data and identifying potential anomalies, which can be crucial in educational research and practice.

Based on the visualizations presented in Figure 5, Figure 6, Figure 7, and Figure 8, the dataset appears to exhibit imbalance, where the data from one class is significantly more prevalent than the other class[30]. This imbalance phenomenon is a common challenge in data analysis, as the disproportionate representation of classes can skew the results and lead to biased conclusions. When dealing with imbalanced datasets, it is crucial to carefully consider and address this issue to ensure accurate and reliable analysis and findings.

Table 2. DNN Model Design

Layer Type	Output Shape	Parameters	Activation	Note
Input	(None, 44)	44 input features	-	The initial layer with 44 features representing student environment variables
Dense	(None, 128)	128 neurons	ReLU	The first hidden layer captures initial complex non-linear relationships
Dropout	(None, 128)	Rate: 0.3	-	Aggressive regularization to prevent overfitting, unusually high dropout rate
Dense	(None, 32)	32 neurons	ReLU	Second hidden layer, further abstract feature extraction
Dense	(None, 32)	32 neurons	LeakyReLU (α=0.02)	Prevents neuron dying problem, allows small gradient for inactive units
Dense	(None, 16)	16 neurons	Softplus	Smooth, continuous activation function for non-linear transformation
Dropout	(None, 16)	Rate: 0.3	-	Additional regularization to enhance model generalization
Dense	(None, 8)	8 neurons	ReLU	Reduces feature dimensionality, extracts more abstract representations
Dropout	(None, 8)	Rate: 0.3	-	Final regularization layer before classification
Dense	(None, 1)	1 neuron	Sigmoid	Binary classification output produces probability between 0 and 1

The neural network model architecture in Table 2 is built using TensorFlow[31]. It starts with an input layer that has 44 features. The data then flows through dense layers with various activation functions and dropout layers for regularization. The first dense layer has 128 neurons with ReLU activation, followed by a dropout layer. The second dense layer has 32 neurons with ReLU activation, followed by LeakyReLU activation, which allows for a small gradient when units are inactive. The third dense layer has 16 neurons with Softplus activation, followed by a dropout layer. The fourth dense layer has 8 neurons with ReLU activation, followed by a dropout layer. Finally, the model has a fifth dense layer with one neuron and Sigmoid activation, which is suitable for binary classification tasks as it produces values between 0 and 1[32].

The accuracy and loss graphs in Figure 9 show the model's performance using the entire dataset without variable selection. The training accuracy increases significantly during the initial epochs, but the validation accuracy appears unstable, experiencing large fluctuations after a few iterations. The loss graph indicates that the training loss value decreases sharply in the beginning but tends to level off in the final epoch, while the validation loss is seen to increase again in the final stage, suggesting overfitting. This overfitting may occur because the model learns too extensively from the training data without adequately considering generalization to the validation data. This aligns with the recommendation about the importance of feature selection to reduce the risk of overfitting[33].

Figure 10 shows the training performance of the model using only the five variables with the highest correlation to the target. The accuracy graph demonstrates more consistent increases in training and validation accuracies, with less fluctuation than the previous model. This trend indicates greater stability, particularly in the validation data, suggesting improved generalization ability. The loss graph also reflects more stable results, where the training loss value decreases significantly without a large gap between the training and validation losses, implying less overfitting than the previous model.

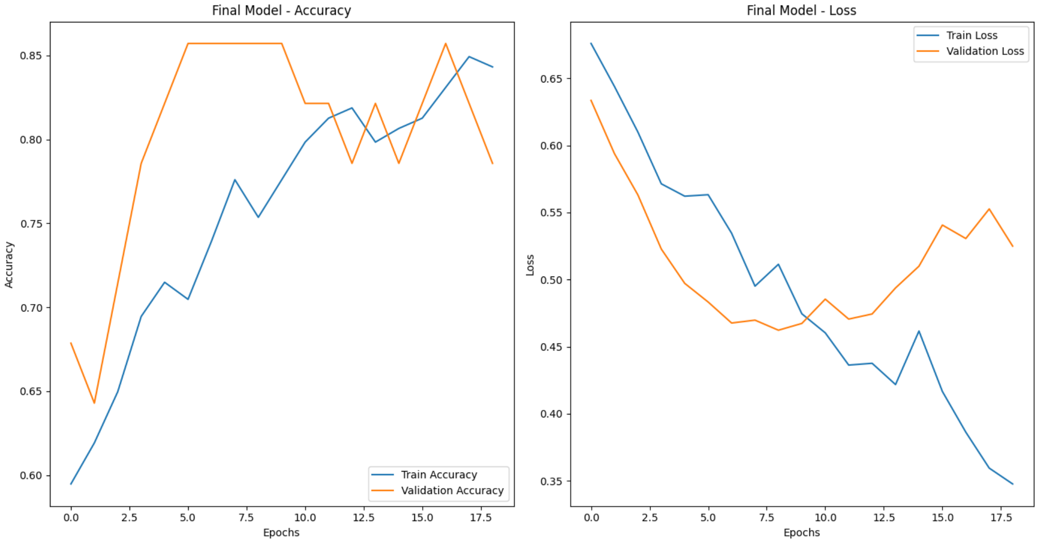

Figure 9. Training Results of Model with All Data

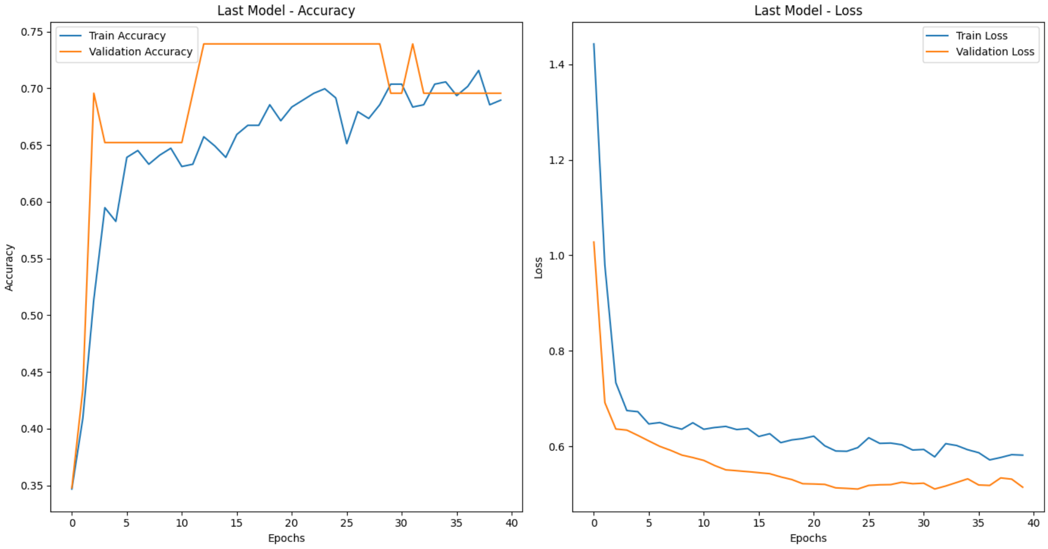

Figure 10. Training Results only Top 5 Correlation

Probabilistic metrics such as Precision, Recall, and F1 can thoroughly evaluate model performance[19]. In this context, the stability observed in Figure 10 can be attributed to using the most relevant variables, which allows the model to optimize the confidence in its predictions. High model confidence in its predictions can also improve the interpretability and quality of the evaluation.

Figure 11 depicts the model training performance when the five variables most correlated with the target are excluded from the data. The accuracy graph exhibits more excellent stability than Figure 10, with both training and validation accuracies demonstrating a consistent upward trend. While the validation accuracy doe s display minor fluctuations in the final epochs, these fluctuations remain within reasonable limits. They are not substantial, suggesting the model maintains a degree of stability despite the omission of important variables. The loss graph indicates that the training and validation loss values decrease significantly in the initial training stages, with a relatively small disparity between them until the midpoint of the training epoch. However, in the final epoch, the validation loss increases while the training loss continues to decline. This observation suggests the model exhibits signs of overfitting toward the end of the training process, as noted in the literature[15], where excluding critical features can amplify the risk of bias and constrain the model's capacity to generalize effectively.

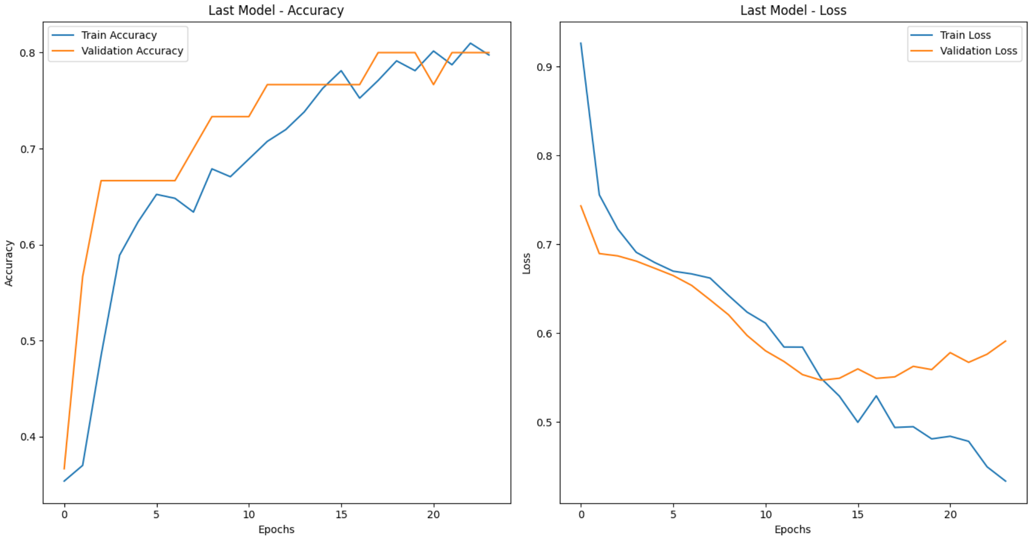

Figure 11. Training Results without Top 5 Correlation

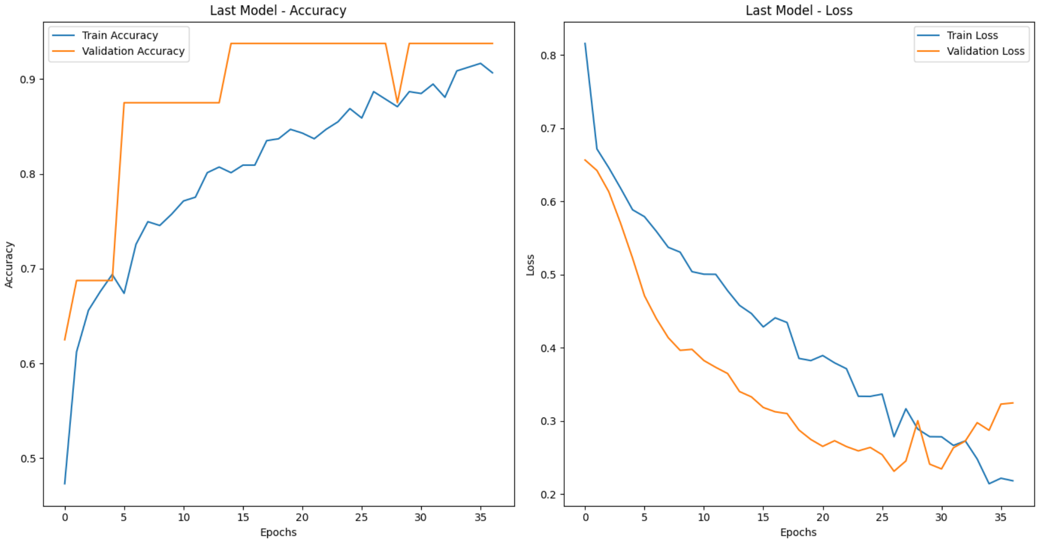

Figure 12. Training Results with Top 5 Correlation

Figure 12 demonstrates that the model trained using the five variables most strongly correlated with the target variable exhibits greater accuracy and loss metrics stability than the model that incorporated all variables without feature selection. On the left side of the graph, the training accuracy shows a gradual upward trend over 35 epochs, while the validation accuracy attains consistently high values. This suggests the model can effectively learn the underlying patterns in the data without exhibiting substantial signs of overfitting, as evidenced by the validation accuracy stabilizing at an elevated level after the initial training stages[15], [19].

Meanwhile, in the graph on the right side, the training loss and validation loss decrease significantly at the beginning of training and start to approach stability at later epochs. Figure 12 demonstrates that the model successfully minimizes the error on both data sets, indicating good generalization to the validation data. By focusing on the most relevant variables, the model achieved more optimal results, which shows that feature selection based on correlation can help improve model performance and efficiency[28], [33].

Utilizing the five variables with the highest correlation was an effective strategy for enhancing the model's accuracy and stability, aligning with the insight that feature selection can assist in reducing model complexity and improving interpretability[34].

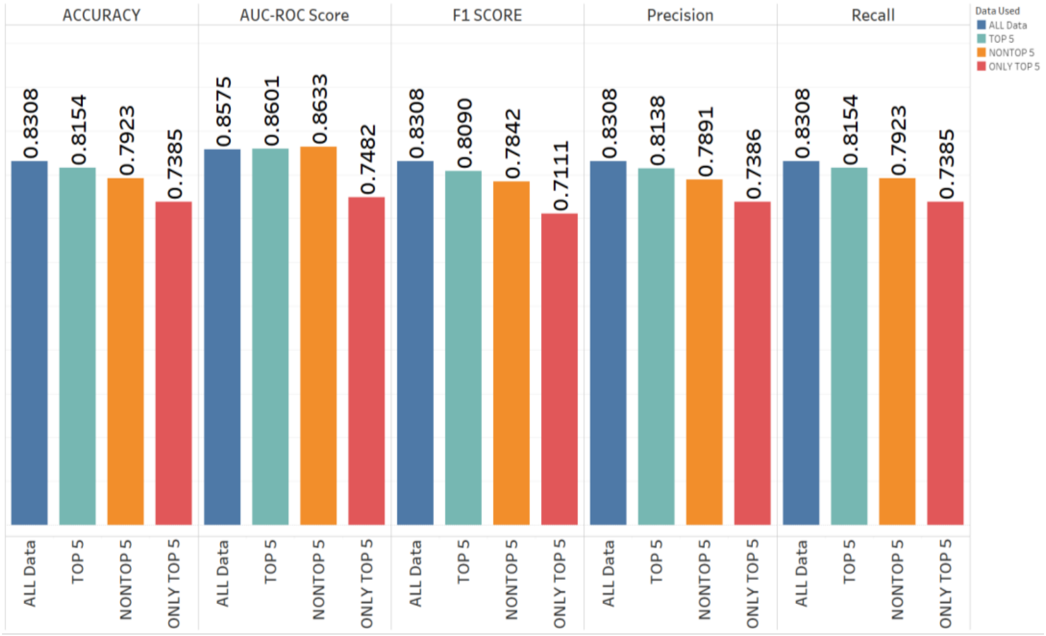

Figure 13. Evaluation Model from All Model

Based on Figure 13 shows a comparison of the various model performance metrics for four variable usage scenarios: All Data, Top 5, Non-Top 5, and Only Top 5. Each of these metrics provides insight into the impact of variable combinations on the classification model's predictive results.

The Accuracy metric demonstrates that the model incorporating all variables, referred to as the "All Data" model, achieved the highest value of 0.8308, indicating it made the most precise predictions. This was followed by the "Top 5" model, which had an accuracy of 0.8154 and performed reasonably well while utilizing a more limited set of variables. The "Non-Top 5" model exhibited a slightly lower accuracy of 0.7923, while the "Only Top 5" model had the lowest accuracy of 0.7385. These findings confirm the significance of highly correlated variables but suggest that additional variables beyond the Top 5 can enhance the model's accuracy, aligning with insights from feature selection research [33], [34].

The Non-Top 5 model demonstrated the highest AUC-ROC Score of 0.8633, narrowly surpassing the performance of the All Data and Top 5 models. This suggests that while the Non-Top 5 variables lack a strong direct correlation with the target, they still contribute meaningful information for distinguishing between the positive and negative classes. Conversely, the Only Top 5 model exhibited the lowest AUC-ROC Score of 0.7482. This corroborates the finding that relying exclusively on highly correlated variables, without incorporating other relevant factors, can diminish the model's capability to detect class variation[15].

The All Data model exhibited the strongest performance in the F1 Score metric, with a value of 0.8308, closely followed by the Top 5 model at 0.8090. The F1 Score, which balances Precision and Recall, indicates that the All Data and Top 5 models achieved more well-rounded results regarding prediction accuracy and range. In contrast, the Non-Top 5 model recorded an F1 Score of 0.7842, while the Only Top 5 model had a lower value of 0.7111. This supports the literature suggesting that feature selection should consider factors beyond correlation to achieve optimal prediction[19].

The Precision values for the All Data and Top 5 models are also relatively high, at 0.8308 and 0.8138, respectively, indicating that these two models are more effective in accurately predicting the positive classes without generating excessive false positives. In contrast, the Non-Top 5 model has a lower Precision value of 0.7891, while the Only Top 5 model exhibits the lowest precision at 0.7386. These findings align with the results of previous studies, which demonstrate that including relevant features can enhance the accuracy of models in detecting the target class[28].

The All Data model achieves the highest Recall value of 0.8308, followed by the Top 5 and Non-Top 5 models. The high Recall results in the All Data and Top 5 models suggest a stronger ability to correctly identify positive instances, with a lower risk of false negative errors. In contrast, the Only Top 5 model recorded the lowest Recall value at 0.7385, indicating that incorporating additional variables can help improve the model's predictive range, in line with findings emphasizing the importance of feature combinations for maintaining model generalizability[19], [33].

These results indicate that using all available variables achieves the best performance across most metrics, while the Top 5 models also perform reasonably well. In contrast, models relying solely on the five most highly correlated variables tend to have the lowest performance, suggesting that additional features not included in the Top 5 still significantly contribute to improving the prediction quality and the model's generalization. This finding supports the literature highlighting the importance of careful feature selection in achieving a balance between the accuracy and interpretability of classification models[19], [34].

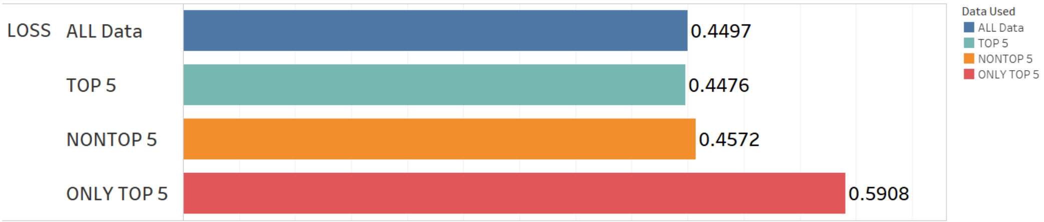

Figure 14. Evaluation Loss Model Each Model

Based on the loss values shown in Figure 14 for various combinations of model variables, there are significant differences in prediction error rates across the four scenarios: All Data, Top 5, Non-Top 5, and Only Top 5. The model utilizing all variables, or the All Data model, produces the lowest loss value of 0.4497, indicating the best performance in terms of error minimization. By leveraging all available information, this model can make more accurate predictions without overlooking important aspects of the data.

The model that uses the five variables with the highest correlation to the target, referred to as the Top 5 model, shows a slightly lower loss value of 0.4476, which is close to the performance of the All Data model. These findings suggest that a model focused on the most relevant variables can significantly reduce complexity while preserving performance. This is consistent with studies demonstrating that effective feature selection can reduce noise and improve model generalization without sacrificing important information[33].

In contrast, the model using the Non-Top 5 variables exhibits a higher loss value of 0.4572. This suggests that variables with low correlation to the target contribute less to improving prediction accuracy and may even introduce uncertainty into the model. This finding emphasizes the importance of selecting variables that significantly influence the target, as supported by previous research on effective feature selection[15].

The model that uses only the five variables with the highest correlation, referred to as the Only Top 5 model, exhibits the highest loss value of 0.5908. This suggests that while the top variables are essential, incorporating additional variables beyond the Top 5 can significantly improve the model's predictive accuracy. Combining highly correlated and other variables provides more optimal results in minimizing the prediction error. This finding underscores the importance of maintaining a balanced approach to feature selection, considering not only the main variables but also the contribution of other variables for better overall performance[19], [34].

Models incorporating all available predictors or a strategic combination of the most salient predictors demonstrate superior performance, as evidenced by lower loss values, compared to models relying on limited or less relevant variables. This suggests that a thoughtful and balanced feature selection strategy is critical for optimizing classification model performance. Such an approach effectively leverages the key predictors while also considering the potential value added by other variables to enhance the overall predictive accuracy [28].

5. Conclusions

This research has successfully analyzed and verified the significant influence of family background and social environment on the scope of schools that students attend, using Deep Neural Networks technology. The results confirm that variables such as family socioeconomic conditions and the local educational environment are important in determining students' access to the quality and type of education.

The findings demonstrate that the developed DNN model can effectively process and analyze complex data to identify patterns and relationships not easily observed through traditional analysis methods. The model's high accuracy and reliability prove its potential for mapping and improving targeted educational interventions. Specifically, the 'All Data' model offers a comprehensive overview of education performance based on various factors, achieving an impressive accuracy of 83.08% and demonstrating its robust analytical capabilities. Meanwhile, the 'Top 5 Data' model effectively shows the influence of the five most essential variables in a more focused manner, with an accuracy of 81.54%. These models' high performance, supported by an AUC-ROC score of 0.8633 and a low loss value of 0.4497, provides valuable insights into the key elements that could be targeted for more efficient and effective educational interventions.

From a practical standpoint, the study's results can help policymakers and educators develop better strategies for resource allocation and student support programs, particularly in socially and economically disadvantaged areas. By identifying and addressing critical factors that affect students' educational opportunities, significant improvements can be made to the quality of education. However, this study has limitations, particularly regarding the limited data coverage of a few schools in a specific region. Combining data from multiple sources and a wider area is recommended for future research to validate these findings and improve the model's generalizability. Furthermore, further exploration of the influence of individual factors such as student academic achievement and learning motivation may provide a deeper understanding of the dynamics of education. This conclusion confirms the great potential of using machine learning technology in education as an analytical tool and a strategic component in education reform and learning policy.

Author Contributions: Conceptualization: M.D.P.K. and A.Z..; methodology, M.D.P.K.; software: M.D.P.K.; validation: M.D.P.K. and A.Z.; formal analysis: M.D.P.K.; investigation: M.D.P.K.; resources: A.Z.; data curation: A.Z.; writing—original draft preparation: M.D.P.K.; writing—review and editing: A.Z; visualization: M.D.P.K.; supervision: A.Z.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] J. Huang, W. Li, and C. Liu, "The Influence of Family Environment on Secondary School Students' Academic Performance," *Lecture Notes in Education Psychology and Public Media*, vol. 12, no. 1, pp. 327–335, Oct. 2023, doi: 10.54254/2753-7048/12/20230837.

[2] Muzamil Hussain AL Hussaini and Tariq Hussain, "Impact of Environmental Factors on Student's Educational Background and Their Future Performance," *International Journal of Scientific Multidisciplinary Research*, vol. 1, no. 6, pp. 607–616, Jul. 2023, doi: 10.55927/ijsmr.v1i6.5021.

[3] W. Yiting, "Study on the influencing factors of family environment on the learning habits of primary school students in middle school," *Applied & Educational Psychology*, vol. 4, no. 11, pp. 6–10, Nov. 2023, doi: 10.23977/appep.2023.041102.

[4] A. Zubair, "Klasifikasi Ketertarikan Mahasiswa Pembelajaran Berbasiskan Video Business Intelligence Menggunakan Adaboost," in *Seminar Nasional Sistem Informasi (SENASIF)*, 2022, pp. 3050–3058.

[5] I. H. Nisa, "The Role Of The School Environment On Student Learning Success, A Descriptive Study at MTS Dzunnuraini," *TARQIYATUNA: Jurnal Pendidikan Agama Islam dan Madrasah Ibtidaiyah*, vol. 1, no. 1, pp. 24–29, Mar. 2022, doi: 10.36769/tarqiyatuna.v1i1.190.

[6] N. Walia, M. Kumar, N. Nayar, and G. Mehta, "Student's Academic Performance Prediction in Academic using Data Mining Techniques," Apr. 01, 2020, *Social Science Research Network, Rochester, NY*: 3565874. doi: 10.2139/ssrn.3565874.

[7] F. Mustofa, A. N. Safriandono, A. R. Muslikh, and D. R. I. M. Setiadi, "Dataset and Feature Analysis for Diabetes Mellitus Classification using Random Forest," *Journal of Computing Theories and Applications*, vol. 1, no. 1, pp. 41–48, Jan. 2023, doi: 10.33633/jcta.v1i1.9190.

[8] S. Mittal, "A survey on modeling and improving reliability of DNN algorithms and accelerators," *Journal of Systems Architecture*, vol. 104, p. 101689, Mar. 2020, doi: 10.1016/j.sysarc.2019.101689.

[9] C. Schröder and A. Niekler, "A Survey of Active Learning for Text Classification using Deep Neural Networks," Aug. 17, 2020, *arXiv*. doi: 10.48550/arXiv.2008.07267.

[10] B. Fu, Q. Shang, T. Sun, and S. Jia, "A Distracted Driving Detection Model Based On Driving Performance," *IEEE Access*, vol. 11, pp. 26624–26636, 2023, doi: 10.1109/ACCESS.2023.3257238.

[11] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.

[12] H. Y. Song and S. Park, "An Analysis of Correlation between Personality and Visiting Place using Spearman's Rank Correlation Coefficient," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 5, May 2020, doi: 10.3837/tiis.2020.05.005.

[13] N. Khatun, "Applications of Normality Test in Statistical Analysis," *Open Journal of Statistics*, vol. 11, no. 01, pp. 113–122, 2021, doi: 10.4236/ojs.2021.111006.

[14] S. Korkmaz and Y. Demir, "Investigation of Some Univariate Normality Tests In Terms of Type-I Errors and Test Power," *Journal of Scientific Reports-A*, no. 052, pp. 376–395, Mar. 2023, doi: 10.59313/jsr-a.1222979.

[15] O. Chamorro-Atalaya *et al.*, "K-Fold Cross-Validation through Identification of the Opinion Classification Algorithm for the Satisfaction of University Students," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 11, Aug. 2023, doi: 10.3991/ijoe.v19i11.39887.

[16] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," Apr. 2008.

[17] D. Marcondes, A. Simonis, and J. Barrera, "Feature Selection based on the Local Lift Dependence Scale," *Entropy*, vol. 20, no. 2, p. 97, Jan. 2018, doi: 10.3390/e20020097.

[18] A. B. F. Mansur and N. Yusof, "The Latent of Student Learning Analytic with K-mean Clustering for Student Behaviour Classification," *Journal of Information Systems Engineering and Business Intelligence*, vol. 4, no. 2, p. 156, Oct. 2018, doi: 10.20473/jisebi.4.2.156-161.

[19] R. Yacouby and D. Axman, "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models," in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 79–91. doi: 10.18653/v1/2020.eval4nlp-1.9.

[20] O. K. Oyedotun, S. N. Tackie, E. O. Olaniyi, and A. Khashman, "Data Mining of Students' Performance: Turkish Students as a Case Study," *International Journal of Intelligent Systems and Applications*, vol. 7, no. 9, pp. 20–27, Sep. 2015, doi: 10.5815/ijisa.2015.09.03.

[21] M. Karagiannopoulos, D. Anyfantis, S. Kotsiantis, and P. Pintelas, "A Wrapper for Reweighting Training Instances for Handling Imbalanced Data Sets," in *Artificial Intelligence and Innovations 2007: from Theory to Applications*, vol. 247, C. Boukis, A. Pnevmatikakis, and L. Polymenakos, Eds., Boston, MA: Springer US, 2007, pp. 29–36. doi: 10.1007/978-0-387-74161-1_4.

[22] N. E. Helwig, "Adding bias to reduce variance in psychological results: A tutorial on penalized regression," *The Quantitative Methods for Psychology*, vol. 13, no. 1, pp. 1–19, Jan. 2017, doi: 10.20982/tqmp.13.1.p001.

[23] The pandas development team, "pandas-dev/pandas: Pandas," Jun. 2023, *Zenodo*. doi: 10.5281/zenodo.8092754.

[24] O. Grisel *et al.*, "scikit-learn/scikit-learn: Scikit-learn 1.5.0," Zenodo. Accessed: Jul. 13, 2024. [Online]. Available: https://zenodo.org/records/11237090

[25] R. Gommers *et al.*, "scipy/scipy: SciPy 1.11.4," Zenodo. Accessed: Jul. 13, 2024. [Online]. Available: https://zenodo.org/records/10155614

[26] The Matplotlib Development Team, "Matplotlib: Visualization with Python," Zenodo. Accessed: Jul. 13, 2024. [Online]. Available: https://zenodo.org/records/11201097

[27] P. Cortez, "Student Performance," UC Irvine Machine Learning Repository. Accessed: Jul. 13, 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/320

[28] W. N. Wassouf, R. Alkhatib, K. Salloum, and S. Balloul, "Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study," *Journal of Big Data*, vol. 7, no. 1, p. 29, Dec. 2020, doi: 10.1186/s40537-020-00290-0.

[29] N. Yau, *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics Second Edition*. Indianapolis: Wiley, 2024. doi: 10.1002/9781394319817.

[30] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed, "Stop Oversampling for Class Imbalance Learning: A Review," *IEEE Access*, vol. 10, pp. 47643–47660, 2022, doi: 10.1109/ACCESS.2022.3169512.

[31] TensorFlow Developers, "TensorFlow," Zenodo. Accessed: Jul. 13, 2024. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.4724125

[32] H. Pratiwi *et al.*, "Sigmoid Activation Function in Selecting the Best Model of Artificial Neural Networks," *Journal of Physics: Conference Series*, vol. 1471, no. 1, p. 012010, Feb. 2020, doi: 10.1088/1742-6596/1471/1/012010.

[33] C. Aliferis and G. Simon, "Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI," in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences*, G. J. Simon and C. Aliferis, Eds., Cham: Springer International Publishing, 2024, pp. 477–524. doi: 10.1007/978-3-031-39355-6_10.

[34] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Machine Learning*, vol. 110, no. 3, pp. 451–456, Mar. 2021, doi: 10.1007/s10994-021-05964-1.