Using AutoML to Predict Student Exam Scores and Analyze Key Performance Factors

Michael Mertl
Data Science
Master Computer Science
Matrikelnr.: 2209076
michael.mertl1@tha.de

***Abstract*—Predicting student performance is a critical aspect of educational research, enabling early identification of at-risk students and the implementation of timely interventions. This study explores the integration of diverse student performance factors to develop accurate predictive models for exam scores. Using the Kaggle "Student Performance Factors" dataset, a systematic machine learning pipeline was implemented, leveraging AutoGluon's automated optimization capabilities and Kedro's modular workflow framework. Key predictors such as attendance and hours studied were identified, with the *WeightedEnsemble_L3* model emerging as the best-performing approach. It achieved an accuracy of 87.77%, Mean Absolute Error (MAE) of 0.2978, and R^2 of 0.9895 (without outliers), while falling short in the Mean Squared Error (MSE) compared to the Ridge Regression model from related studies. These findings underscore the trade-offs between optimizing different evaluation metrics and highlight the effectiveness of ensemble-based learning for improving accuracy and explained variance, while emphasizing practical implications for educators and policymakers by identifying critical factors like attendance and hours studied to guide targeted interventions. Despite the limitations of using a synthesized dataset, this research demonstrates the scalability and efficiency of automated machine learning frameworks in advancing educational analytics.**

***Index Terms*—Student performance prediction, automated machine learning, ensemble models, AutoGluon, educational data mining, predictive analytics, Kaggle dataset, Kedro.**

I. Introduction

Predicting student performance is a critical area of educational research, as it enables early identification of at-risk students and facilitates timely interventions. By leveraging data generated from students' participation in academic activities, such as peer-assessment tasks, educators can gain insights into students' engagement and performance trends. Studies have demonstrated that prediction models using such data are valuable for assessing academic success and informing interventions to improve student outcomes. This approach aligns with the broader goal of utilizing student-generated data to provide continuous assessment and supervision in educational settings [1].

Accurately predicting exam scores remains a considerable challenge in educational research, despite its importance for improving academic outcomes. Traditional predictive approaches, which often rely on standardized test scores or prior academic achievements, fail to capture the multifaceted nature of modern learning environments, including behavioral and contextual factors [2]. Crucial variables such as attendance, study habits, and participation in extracurricular activities are frequently overlooked, even though they exhibit a strong correlation with academic performance [2].

The widespread adoption of online learning platforms has further amplified the complexity of this task. These platforms generate extensive data from student interactions, participation, and assessments, requiring new methodologies to harness this wealth of information effectively [3]. A major obstacle lies in integrating and analyzing diverse data types, such as numerical scores, categorical variables, and behavioral metrics. This necessitates robust computational frameworks capable of managing heterogeneous datasets while uncovering meaningful patterns [2], [3].

As educational institutions increasingly collect detailed data on student behavior and performance, the demand for sophisticated predictive models that leverage this information has become more pressing. These models must bridge the gap between traditional approaches and the dynamic, data-rich environments of contemporary education [2], [3].

This research seeks to address these challenges by investigating the prediction of exam scores through the integration of diverse student performance factors. By employing advanced data analysis methodologies and automated machine learning techniques, this study aims to identify the most influential determinants of academic success and develop highly accurate predictive models. The findings of this work aspire to contribute to the broader objectives of improving student outcomes and advancing the field of educational research.

This report adopts a slightly unconventional structure to better align with its objectives. Following the *Introduction*, the *Related Work* section is presented first to provide a contextual overview of prior research efforts and existing benchmarks in predicting student performance. This establishes a foundation for the subsequent *State of the Art* section, which describes the advanced technologies and frameworks utilized in this study—namely, AutoGluon and Kedro. The final research question is then introduced, synthesizing the challenges, related efforts, and the capabilities of the tools described to set the stage for the methodology and analysis that follow.

II. Related Work

The prediction of students' academic performance has been a topic of significant interest in educational data mining. Various models and datasets have been utilized in prior research to address similar challenges, with diverse methodologies and outcomes.

A. Studies on different datasets

The work by Mogessie et al. [1] presents a linear regression model to predict final exam scores based on students' engagement in peer-assessment activities. The data was gathered from two undergraduate courses, using features such as tasks completed, votes earned, and perceived difficulty of questions. The model achieved a Root Mean Squared Error (RMSE) of 2.93 on one course and 3.44 on another with a scoring scale of 18 to 30, demonstrating its robustness across varying contexts. This study underscores the value of continuous activity tracking and student engagement as predictive factors.

In another study, Sokkhey and Okazaki [4] developed a web-based system for predicting poor-performing students using educational data mining techniques. This research utilized a dataset of student interactions, assessments, and participation metrics. The authors introduced a hybrid Random Forest model combined with a MICHI feature selection method, which integrates mutual information and chi-square algorithms to identify key factors influencing academic performance. The model was evaluated on a classification task to predict at-risk students and achieved an exceptional accuracy of 99.98% and an RMSE of 0.008. This work highlights the importance of advanced machine learning techniques and rigorous feature selection in building highly accurate academic performance prediction systems.

B. Studies on the same dataset

The Kaggle dataset "Student Performance Factors" [5], which forms the foundation of this research and is described in detail in section V.A ("Dataset"), has been extensively explored in previous studies. Asadozzaman [6] evaluated various models and reported a Mean Squared Error (MSE) of 3.7384 and a Mean Absolute Error (MAE) of 0.7440 using a Support Vector Regressor (SVR) as the best-performing model. Similarly, Ibrahem [7] compared multiple regression models and identified a Ridge Regression model as the best approach, with an MSE of 2.9145 and an accuracy of 79.70%. Rafazi [8] developed a neural network to predict exam scores, achieving an MAE of 0.8646. Gamal [9] applied a linear regression model but excluded outliers, resulting in an R-squared (R^2) score of 0.9513. These studies underscore the dataset's utility and establish important benchmarks for predictive performance, against which the current research can be compared.

C. Comparison with this research

Building on the contributions of previous studies utilizing the Kaggle dataset, this research seeks to further advance the understanding of student performance by integrating state-of-the-art machine learning tools and systematic evaluation techniques. Earlier works have established important benchmarks for predictive performance, identifying both the potential and limitations of various modeling approaches.

This study expands on these efforts by employing advanced frameworks to analyze the dataset and identify the most influential factors affecting academic outcomes. The methodologies and findings of this research are systematically compared with those from related studies to assess improvements in predictive accuracy. The details of this comparison are presented in section VI.F ("Comparison with related Kaggle work"), following a comprehensive discussion of the employed tools and research methodology in the upcoming sections.

III. State of the Art

The development of predictive models for student performance requires robust and scalable frameworks that can handle diverse data types and automate complex machine learning tasks. This study leverages two state-of-the-art technologies: AutoGluon and Kedro. These tools streamline model development and enhance the reproducibility of the research.

A. AutoGluon

AutoGluon[1] is a powerful automated machine learning (AutoML) framework designed to optimize model performance with minimal manual intervention. By automating key steps such as feature selection, hyperparameter tuning, and model evaluation, AutoGluon simplifies the machine learning pipeline, making it highly efficient for various predictive tasks [10].

1) TabularPredictor: The *TabularPredictor* is one of AutoGluon's most notable components, specifically designed for tabular data. It supports regression and classification tasks, offering a range of models including tree-based methods, neural networks, and ensemble techniques [10], [11]. The *TabularPredictor* automatically [10]:

- Handles missing values and categorical features through encoding.
- Evaluates multiple models with different configurations using its built-in presets, such as `best_quality` for high accuracy and `fastest` for rapid inference.
- Generates a leaderboard showcasing the performance of all tested models.

This feature-rich module ensures that the best model for a given dataset is selected with minimal computational effort.

2) MultiModalPredictor: The *MultiModalPredictor* extends AutoGluon's capabilities by integrating diverse data types such as text, images, and tabular data into a single predictive model. This approach is particularly advantageous for datasets that combine numerical attributes with unstructured data, enabling holistic analysis and seamless multimodal integration [12], [13].

[1]For more information on AutoGluon, refer to its official documentation: https://auto.gluon.ai/stable/index.html.

B. Kedro

Kedro[2] is an open-source Python framework for creating modular, reproducible, and maintainable data science workflows. It provides a structured pipeline architecture, allowing researchers to focus on the logic of data processing and model development without being burdened by operational complexities [14].

Key benefits of Kedro include [14].:

- *Modularity*: Each stage of the data science workflow is treated as an independent component, ensuring flexibility and reusability.
- *Version Control*: Kedro integrates with version control systems to track changes in datasets, models, and code, ensuring reproducibility.
- *Pipeline Visualization*: A built-in visualization tool provides an intuitive overview of the data pipeline, facilitating debugging and optimization.

By integrating Kedro with AutoGluon, this study establishes a robust foundation for handling the dataset, preprocessing, model training, and evaluation in a structured and efficient manner.

IV. Research Question

Building on the challenges and methodologies outlined in the previous sections, this study focuses on uncovering the most significant factors that influence student performance and leveraging these insights to develop accurate predictive models. While prior research has demonstrated the utility of various datasets and machine learning techniques, as discussed in section II (”Related Work”), and state-of-the-art tools such as AutoGluon and Kedro offer powerful capabilities for automating and structuring workflows, there remains a need to synthesize these advancements to address specific educational contexts.

To address this gap, this study investigates the following research question:

Which factors influence student performance the most, and how can we predict a student's performance based on these factors?

By integrating the methodologies and tools discussed earlier, this research aims to identify key predictors of academic success while evaluating the effectiveness of automated machine learning frameworks and structured pipelines.

V. Methodology

To address the research question outlined in this study, a structured and systematic methodological framework was designed. This framework focuses on the development of a machine learning pipeline leveraging Kedro and multiple AutoGluon approaches to identify the optimal predictive model. The subsequent sections provide a detailed exposition of the methodologies and techniques employed, emphasizing their relevance and alignment with the research objectives.

[2]For more information on Kedro, refer to its official documentation: https://docs.kedro.org/en/stable/introduction/index.html.

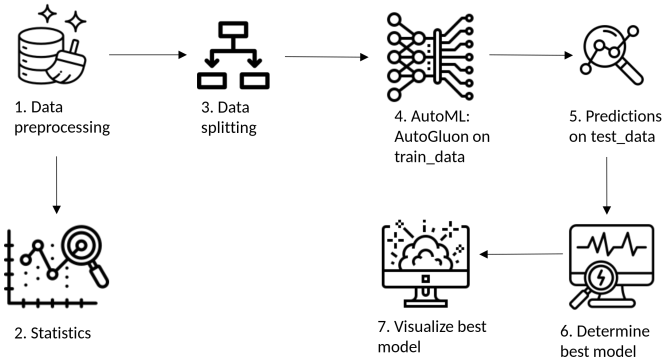

Fig. 1. Overview of the Machine Learning Pipeline Architecture

A. Dataset

The dataset utilized for this research, titled ”Student Performance Factors,” was sourced from Kaggle [5]. This dataset offers a comprehensive analysis of various elements influencing student performance in examinations. It encompasses a diverse range of attributes, including hour studied habits, attendance, parental involvement, and other factors critical to academic success. Table I presents an overview of the dataset's columns and attribute types.

The dataset is characterized by a Kaggle usability score of 10, indicating its exceptional completeness, credibility, and compatibility for analytical tasks. It comprises 6,607 entries across 20 features. As of the last verification on the 3rd of January, 2025, the dataset had garnered 235,000 views, 59,200 downloads, and was cited in 69 research studies, underscoring its widespread acceptance and reliability.

It is important to note that the dataset is synthesized, which may limit its applicability to real-world educational settings. This limitation is acknowledged and discussed in detail in section VII (”Discussion”), as it underscores the need for further validation on real-world datasets to ensure the robustness and generalizability of the findings.

B. Pipeline Architecture

The machine learning pipeline designed for this study is composed of seven key stages, as illustrated in Figure 1. The pipeline leverages Kedro for its modular and reproducible workflow capabilities and integrates AutoGluon for automated machine learning (AutoML). Each stage is described in detail below:

1) **Data Preprocessing:** The pipeline begins with the preprocessing stage, where the dataset is loaded and inspected for missing values. Columns with missing values are identified and removed to ensure the dataset's integrity and compatibility with subsequent stages.
2) **Statistical Analysis:** The second stage focuses on statistical exploration of the dataset, which contributes to identifying key factors influencing the target variable, *Exam Score*. The following analyses are performed:

TABLE I
COLUMNS DESCRIPTION OF THE DATASET FROM KAGGLE [5]

Attribute	Description
Hours_Studied	Number of hours spent studying per week.
Attendance	Percentage of classes attended.
Parental_Involvement	Level of parental involvement in the student's education (Low, Medium, High).
Access_to_Resources	Availability of educational resources (Low, Medium, High).
Extracurricular_Activities	Participation in extracurricular activities (Yes, No).
Sleep_Hours	Average number of hours of sleep per night.
Previous_Scores	Scores from previous exams.
Motivation_Level	Student's level of motivation (Low, Medium, High).
Internet_Access	Availability of internet access (Yes, No).
Tutoring_Sessions	Number of tutoring sessions attended per month.
Family_Income	Family income level (Low, Medium, High).
Teacher_Quality	Quality of the teachers (Low, Medium, High).
School_Type	Type of school attended (Public, Private).
Peer_Influence	Influence of peers on academic performance (Positive, Neutral, Negative).
Physical_Activity	Average number of hours of physical activity per week.
Learning_Disabilities	Presence of learning disabilities (Yes, No).
Parental_Education_Level	Highest education level of parents (High School, College, Postgraduate).
Distance_from_Home	Distance from home to school (Near, Moderate, Far).
Gender	Gender of the student (Male, Female).
Exam_Score	Final exam score.

- A correlation heatmap is generated for all numerical attributes in the dataset to visualize relationships between features.
- Two scatter plots are created to examine the strongest correlations with the target variable.
- A custom encoding map is implemented to convert categorical and boolean values into numeric representations, facilitating advanced analysis.
- A comprehensive correlation heatmap is generated for the encoded dataset, providing additional insights into feature relationships.

3) **Data Preparation:** In this stage, the preprocessed dataset is split into two subsets: 90% for training (train set) and 10% for testing (test set). This stratified split ensures that both subsets retain representative distributions of the target variable.
4) **Automated Machine Learning with AutoGluon:** The training data is utilized in three different AutoGluon approaches to identify the optimal predictive model:
   - **TabularPredictor:** The first approach employs the standard TabularPredictor module from AutoGluon with a training time of four hours and the `best_quality` preset. This configuration evaluates multiple models and automatically selects the best-performing one.
   - **TabularPredictor with Neural Networks:** The second approach is identical to the first, but it exclusively trains and evaluates neural network models to identify the best deep learning-based solution.
   - **MultiModalPredictor:** The third approach uses AutoGluon's MultiModalPredictor module, which creates a multimodal model architecture tailored to the dataset. This approach also runs for four hours using the `best_quality` preset.
5) **Model Evaluation:** After training, the best model from each of the three approaches is applied to the test set. Accuracy metrics are computed for each model to assess their predictive performance on unseen data.
6) **Model Comparison:** The computed accuracy scores for the best models from each approach are compared to determine the most effective AutoML AutoGluon method and the corresponding best model.
7) **Visualization of Results:** In the final stage, the leaderboard of the best-performing approach is visualized to display the performance of all models evaluated by that approach. The top model is further analyzed and visualized, including its architecture and feature importance scores. These feature importance scores are compared with the initial correlation heatmap to verify consistency in feature relevance.

This pipeline ensures a systematic and efficient approach to identify the best predictive model using AutoGluon, while providing clear visualizations and insights for comparison with existing research. By combining data preprocessing, statistical exploration, automated model training, and robust evaluation, the pipeline delivers a comprehensive framework for addressing the research objectives.

C. Comparison with Related Kaggle Work

To evaluate the effectiveness of the best predictive model obtained from the pipeline, its performance is compared with results from related studies. This comparison aims to contextualize the findings of this research within the broader academic literature and determine whether the proposed methodology achieves better, comparable, or worse performance.

The comparison process involves:

1) **Selection of Benchmark Studies:** Key studies addressing similar research questions and using the same datasets are identified from Kaggle (see Section II.B ("Studies on the same dataset")).
2) **Performance Metrics:** Metrics reported in related studies, such as accuracy, MAE or MSE, are extracted to serve as benchmarks. The same metrics calculated for the best model from this study are used to ensure a consistent basis for comparison.
3) **Analysis of Results:** The performance of the best model from this pipeline is assessed relative to the benchmark results to determine its competitiveness and highlight potential areas for improvement.

This comparison process ensures that the contributions of this study are positioned within the context of existing research and provides a clear basis for evaluating the effectiveness of the proposed pipeline.

VI. Results

This section presents the results obtained from the proposed machine learning pipeline, including statistical analyses, model performance metrics, and a comparison with related work. Key visualizations and insights are provided to highlight the effectiveness of the pipeline in identifying the best predictive model and answering the research questions[3].

A. Statistical Analysis Results

The statistical analysis involved multiple visualizations and techniques to explore relationships between features and the target variable, *Exam Score*. This subsection is divided into four parts: two correlation heatmaps and two scatter plots, each focusing on specific results.

1) Correlation Heatmap (Numerical Features): The initial correlation heatmap, shown in Figure 2, illustrates the relationships among numerical features in the dataset. The features *Hours Studied* and *Attendance* demonstrated the strongest positive correlations with the target variable, *Exam Score*, with correlation coefficients of 0.45 and 0.58, respectively. These findings suggest that students who study more hours per week and have higher previous exam scores are more likely to achieve better performance.

[3]For detailed information about the results, the developed pipeline, the conducted research, and the associated code, please visit the project repository at https://github.com/Murtl/data-science.

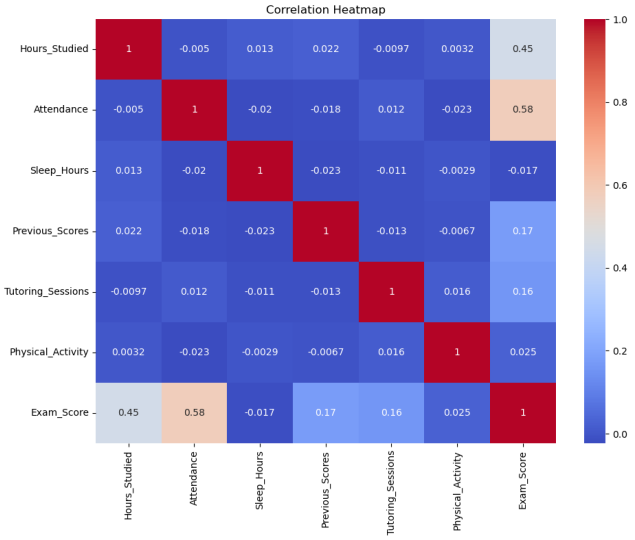

Fig. 2. Correlation heatmap showing relationships between numerical features

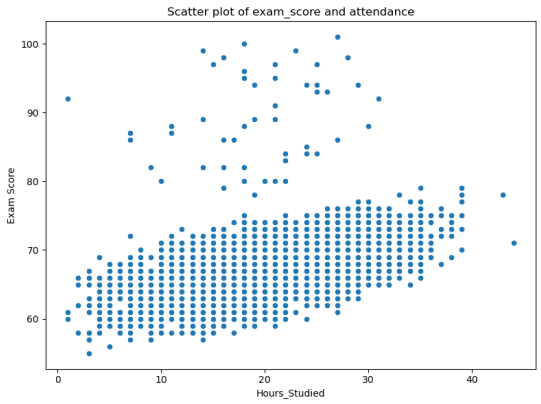

Fig. 3. Scatter plot of *Hours Studied* and *Exam Score*

2) Scatter Plot: Hours Studied and Exam Score: The scatter plot in Figure 3 provides a deeper understanding of the relationship between *Hours Studied* and *Exam Score*. A positive trend is evident, indicating that as students invest more time in studying, their exam scores generally improve. However, notable outliers suggest other influencing factors may moderate this relationship.

3) Scatter Plot: Attendance and Exam Score: Figure 4 highlights the relationship between *Attendance* and *Exam Score*. A similar positive trend is observed, with higher attendance percentages correlating with improved exam performance. The data suggests that regular class attendance plays a critical role in achieving higher academic outcomes.

4) Correlation Heatmap (Encoded Categorical and Boolean Features): A second heatmap, shown in Figure 5, was generated after applying custom encoding to categorical and boolean features. This visualization aimed to identify additional relationships between encoded variables and *Exam Score*. However, no new significant correlations or insights

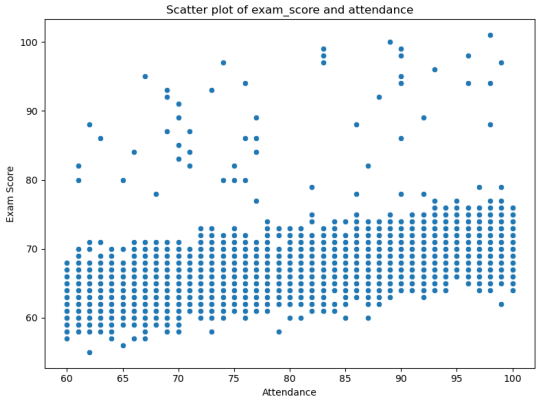

Fig. 4. Scatter plot of *Attendance* and *Exam Score*

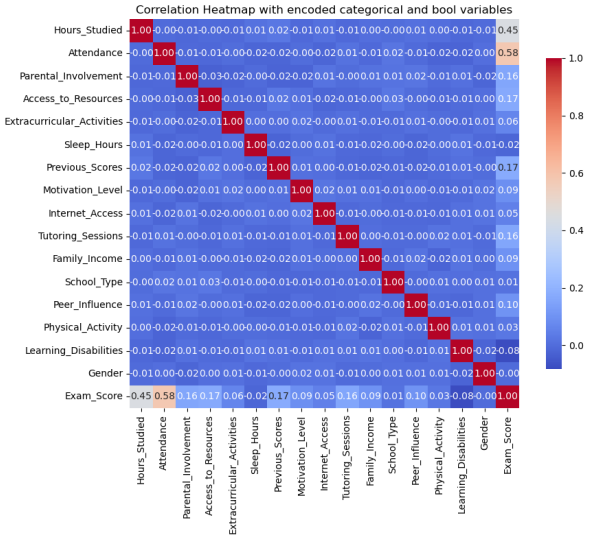

Fig. 5. Correlation heatmap with encoded categorical and boolean variables

were uncovered compared to the initial heatmap of numerical features. Moreover, since AutoGluon inherently performs feature encoding during training, the added complexity of manual encoding did not contribute to meaningful advances in the analysis.

The findings from this encoded heatmap reaffirm that the features *Hours Studied* and *Attendance*, as observed in the initial correlation heatmap, remain the most influential variables for predicting *Exam Score*.

TABLE II
PERFORMANCE COMPARISON OF THE BEST MODELS FROM EACH AUTOGLUON APPROACH

Approach	**Accuracy (%)**
TabularPredictor	**87.777**
TabularPredictor (Neural Networks)	83.072
MultiModalPredictor	81.975

TABLE III
TOP 5 MODELS FROM THE LEADERBOARD (ACCURACY IN %)

Model	**Validation Accuracy (%)**
WeightedEnsemble_L3	87.133
LightGBM_BAG_L2	87.133
RandomForestGini_BAG_L2	86.851
CatBoost_BAG_L2	86.763
RandomForestEntr_BAG_L2	86.569

B. Model Performance

Building upon the insights from the statistical analysis, the next phase of this study focused on evaluating the predictive performance of different AutoGluon approaches. By leveraging the relationships observed in the statistical analysis, the machine learning pipeline was employed to systematically train, validate, and compare predictive models.

Table II summarizes the accuracy scores for the best models obtained from each AutoGluon approach after an extra prediction run on test data. Among the three approaches, the *TabularPredictor* achieved the highest accuracy with its best trained model, indicating its effectiveness in leveraging diverse feature types to predict *Exam Scores*. This finding marks a significant step toward identifying the optimal methodology for addressing the research questions posed.

The best-performing model is further analyzed to understand its architecture and strengths.

C. Leaderboard of Models

Table III displays the leaderboard generated by the *TabularPredictor* approach, showcasing the performance of the top models evaluated. The *WeightedEnsemble_L3* model achieved the highest accuracy and overall position in the leaderboard making it the best-performing model for this study.

It is important to notice, that the leaderboard accuracy is from the validation data that AutoGluon split from the train data. The accuracies in table II were created on the extra test data that is why these values differ.

D. Detailed Analysis of the Best-Performing Model: WeightedEnsemble_L3

The best-performing model identified by the *TabularPredictor* approach is the *WeightedEnsemble_L3* model, which achieved an validation accuracy of 87.133% and an accuracy of 87.777% on the test set. This section provides a detailed analysis of its architecture, functionality, and the features contributing to its performance.

1) Model Overview: The *WeightedEnsemble_L3* model is a weighted ensemble model, which combines predictions from multiple base models to enhance overall predictive accuracy. Ensemble models are particularly effective in leveraging the strengths of individual base models while mitigating their weaknesses [15]. In this study, the ensemble aggregates predic-

TABLE IV
CALCULATED FEATURE IMPORTANCE SCORES FROM THE *WeightedEnsemble_L3* MODEL

Feature	Importance
Attendance	0.756260
Hours_Studied	0.734707
Previous_Scores	0.538510
Parental_Involvement	0.529002
Access_to_Resources	0.502377
Tutoring_Sessions	0.487650

tions from various versions of *LightGBM_BAG_L2*, a bagged version of the LightGBM algorithm[4].

The key characteristics of the *WeightedEnsemble_L3* model are as follows:

- **Model Type:** Weighted Ensemble Model
- **Problem Type:** Multiclass classification
- **Evaluation Metric:** Accuracy
- **Number of Base Models:** 25, with a maximum of 5 base models per type
- **Features Used:** Derived from base models of *LightGBM_BAG_L2*, ensuring diversity in predictive signals

2) How WeightedEnsemble_L3 works: The *WeightedEnsemble_L3* uses a greedy weighted ensemble strategy, where the weights assigned to base models are optimized to minimize classification error [16]. By assigning higher weights to better-performing base models, the ensemble maximizes its overall predictive accuracy. This approach is particularly useful for datasets with diverse feature distributions, as it captures complementary strengths of different models.

Key aspects of the ensemble strategy include:

- **Greedy Weight Optimization:** The ensemble iteratively adjusts weights to optimize performance metrics on the validation set [17].
- **Diversity of Base Models:** By aggregating outputs from multiple LightGBM variants, the ensemble enhances robustness to overfitting [15].
- **Reduced Bias:** Combining models helps to mitigate the biases inherent in individual base models [15].

E. Feature Importance Analysis

The feature importance analysis provides a deeper understanding of how individual features contribute to the predictive performance of the *WeightedEnsemble_L3* model. These importance scores[5] were computed based on the model's ability to explain variance in the target variable, *Exam Score*, during training. The exact scores of the top six features are summarized in Table IV and all scores are visualized in Figure 6.

[4]For more information about LightGBM, refer to the official documentation here: https://lightgbm.readthedocs.io/en/latest/.

[5]For more information about the calculation of the importance scores, refer to https://auto.gluon.ai/stable/api/autogluon.tabular.TabularPredictor.feature_importance.html.

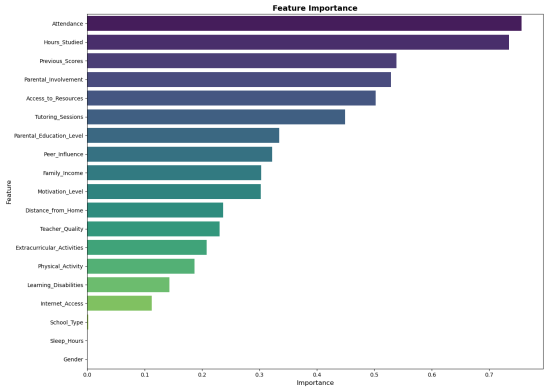

Fig. 6. Feature importance scores from the *WeightedEnsemble_L3* model

The top six features contributing to the model's predictions include:

- **Attendance** (0.756): The highest-ranked feature, emphasizing the critical role of class participation in academic success.
- **Hours Studied** (0.735): A closely ranked feature, reaffirming the importance of consistent study habits.
- **Previous Scores** (0.539): Reflecting prior academic performance, this feature underscores the value of historical data in predicting future outcomes.
- **Parental Involvement** (0.529): Highlighting the influence of parental engagement on a student's performance.
- **Access to Resources** (0.502): Demonstrating the significant impact of educational materials and tools available to students.
- **Tutoring Sessions** (0.488): Confirming the supportive role of additional learning assistance in boosting exam scores.

Figure 6 visualizes the feature importance scores for all features, providing a comprehensive perspective on their relative contributions. These results align with earlier findings, where features such as *Hours Studied* and *Attendance* were strongly correlated with *Exam Score*. The inclusion of other key features, such as *Parental Involvement* and *Access to Resources*, adds valuable context to the interpretation of student performance predictors.

In summary, the feature importance analysis not only validates the significance of previously identified features but also highlights additional influential factors, offering a holistic understanding of the predictors driving student exam performance.

F. Comparison with Related Kaggle Work

To evaluate the effectiveness of the proposed *WeightedEnsemble_L3* model, its performance is compared with other research efforts that utilized the same Kaggle dataset, as described in section II.B ("Studies on the Same Dataset"). This comparison involves metrics such as Accuracy, Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2)—calculated in a consistent manner to ensure

TABLE V
COMPARISON OF MODEL PERFORMANCE WITH RELATED STUDIES.

Model	Accuracy	MSE	MAE	R^2 (w/o outliers)
Support Vector Regressor (Asadozzaman [6])	-	3.7384	0.7440	-
Neural Network (Rafazi [8])	-	-	0.8646	-
Ridge Regression model (Ibrahem [7])	79.70%	**2.9145**	-	-
Linear Regression model (Gamal [9])	-	-	-	0.9513
WeightedEnsemble_L3 (this study)	**87.77%**	4.4514	**0.2978**	**0.9895**

comparability[6]. Table V summarizes the results from related studies alongside the outcomes of this research.

The results presented in Table V showcase the strengths of the *WeightedEnsemble_L3* model across various metrics, though some limitations are evident in certain cases.

Compared to Asadozzaman's Support Vector Regressor (SVR), which achieved an MSE of 3.7384 and an MAE of 0.7440, the *WeightedEnsemble_L3* demonstrates a substantial improvement in MAE, achieving 0.2978—a reduction of approximately 60%. However, its MSE (4.4514) is slightly higher, indicating that while the ensemble model is better at minimizing average absolute errors, it is less effective in reducing squared errors.

Rafazi's Neural Network model reported an MAE of 0.8646 but did not include values for MSE or R^2, making a complete comparison challenging. Still, the *WeightedEnsemble_L3* outperforms in MAE with a 65% reduction, highlighting the ensemble model's strength even against deep learning approaches.

Ibrahem's Ridge Regression model achieved a strong MSE of 2.9145 and an accuracy of 79.70%, showcasing its efficiency among traditional regression techniques. While the *WeightedEnsemble_L3* improves accuracy to 87.77%, it lags behind in MSE, which is 53% higher. This result suggests that the Ridge Regression model excels in minimizing squared prediction errors, whereas the ensemble approach emphasizes accuracy and MAE.

Gamal's Linear Regression model reported an R^2 score of 0.9513 after removing outliers. The *WeightedEnsemble_L3* surpasses this with an R^2 score of 0.9895, if the outliers are removed on the same way as Gamal did it. This shows the models superior ability to explain variance in the target variable.

Overall, while the *WeightedEnsemble_L3* outperforms all other models in terms of accuracy, MAE, and R^2, it does not achieve the lowest MSE. These results highlight the model's strengths in reducing average absolute errors and explaining variance but indicate some trade-offs in its ability to minimize squared errors. This trade-off may result from the ensemble's optimization process prioritizing accuracy and other metrics over strict minimization of MSE.

G. Summary of Findings

The results highlight the strengths and limitations of the proposed machine learning pipeline in identifying a predictive model for student performance. The *WeightedEnsemble_L3* of the *TabularPredictor* approach achieved superior performance in key metrics such as accuracy, Mean Absolute Error (MAE), and R^2, outperforming related studies on the same dataset. However, the *WeightedEnsemble_L3* did not achieve the lowest Mean Squared Error (MSE), with Ibrahem's Ridge Regression model demonstrating a better ability to minimize squared errors.

This nuanced performance underscores the trade-offs inherent in optimizing for different evaluation metrics. Furthermore, the analysis reaffirmed the importance of key features such as *Hours Studied* and *Attendance* in predicting *Exam Scores*, solidifying their role as critical factors in student performance prediction.

VII. DISCUSSION

The findings of this study highlight the effectiveness of using an automated machine learning pipeline to predict student performance and identify key factors influencing exam scores. The proposed *WeightedEnsemble_L3* model, built using AutoGluon's *TabularPredictor*, achieved superior performance metrics, including an accuracy of 87.77%, MAE of 0.2978, and R^2 of 0.9895. However, it did not achieve the lowest Mean Squared Error (MSE), as Ibrahem's Ridge Regression model demonstrated better performance in this metric with an MSE of 2.9145.

This discrepancy highlights the inherent trade-offs between optimizing for different evaluation metrics. Ensemble models like the *WeightedEnsemble_L3* prioritize reducing overall prediction error and maximizing accuracy by aggregating diverse model outputs. However, this approach can lead to slightly higher squared errors when individual predictions deviate significantly from the true values. Such trade-offs suggest that while ensemble methods excel in improving overall robustness, models like Ridge Regression may remain more suitable for specific tasks where minimizing squared errors is critical.

Despite these accomplishments, several limitations of the study should be acknowledged:

1) **Synthetic Dataset:** The Kaggle dataset used in this research is synthesized and may not fully represent real-world scenarios. Although it provides valuable insights, the generalizability of the findings to actual educational settings remains uncertain. This limitation is critical, as real-world data may exhibit greater variability and noise.
2) **Scalability Across Educational Contexts:** The dataset focuses on a generalized context without accounting for regional, cultural, or institutional differences in education. These contextual variances could affect the transferability of the findings to other educational settings.

To address these limitations, future research could explore the following directions:

1) **Real-World Dataset Collection:** Building upon the insights gained from this research, future studies could collect real-world data that include the identified key factors, such as attendance, hours studied, and parental

[6]If you need more information about these metrics, refer to Verma's article, "Understanding Regression Metrics: A Comprehensive Guide" [18].

involvement. Applying the developed pipeline to such datasets would validate its effectiveness and generalizability.

2) **Localized and Contextualized Models:** To enhance scalability, future research should develop models tailored to specific educational contexts. This could involve training region-specific models or incorporating cultural and institutional factors into the dataset and analysis.

Overall, the proposed pipeline offers a robust foundation for advancing predictive analytics in education, but further work is required to enhance its applicability and impact in real-world scenarios.

VIII. Conclusion

This study demonstrated the potential of using an automated machine learning pipeline, powered by AutoGluon and Kedro, to predict student exam scores and identify the most influential performance factors. By analyzing a synthesized Kaggle dataset, the research identified *Attendance* and *Hours Studied* as critical predictors of academic success.

The *WeightedEnsemble_L3* model outperformed other approaches applied to the same dataset in key metrics such as accuracy (87.77%), MAE (0.2978), and R^2 (0.9895, without outliers). However, it did not achieve the lowest Mean Squared Error (MSE), as Ibrahem's Ridge Regression model demonstrated superior performance in this metric. These results highlight the trade-offs between optimizing for different evaluation metrics while establishing the *WeightedEnsemble_L3* as a robust model for accuracy and explained variance.

The results also have practical implications for educators and policymakers. By emphasizing critical predictors like attendance and hours studied, the study provides actionable insights for designing targeted interventions aimed at improving student performance. Schools and institutions can leverage these findings to focus resources on improving class attendance and fostering effective study habits among students.

The study's limitations, including the use of a synthesized dataset and the lack of interaction effects between predictors, emphasize the need for future research to validate these findings in real-world contexts. Additionally, developing localized models that account for contextual and cultural differences in education would enhance the robustness and generalizability of predictive frameworks.

In conclusion, this research advances the understanding of student performance predictors while showcasing the utility of automated machine learning frameworks in addressing complex educational challenges. The proposed methodology offers a scalable and efficient approach that can be extended to diverse datasets and educational contexts, paving the way for more personalized and effective interventions aimed at improving student outcomes.

Acknowledgment

The author expresses sincere gratitude to the Kaggle community for generously providing the "Student Performance Factors" dataset, which served as the cornerstone for this research. Deep appreciation is also extended to academic mentors and peers whose constructive feedback and insightful guidance greatly contributed to the depth and rigor of this study. Furthermore, the author acknowledges the utilization of GPT-4o for its role in refining the manuscript's language, ensuring precision in grammar, clarity of expression, and stylistic cohesion, while maintaining the integrity and scientific rigor of the work.

References

[1] M. M. Ashenafi, G. Riccardi, and M. Ronchetti, "Predicting students' final exam scores from their course activities," in *2015 IEEE Frontiers in Education Conference (FIE)*, 2015, pp. 1–9. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7344081

[2] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Applied Sciences*, vol. 10, no. 3, p. 1042, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/3/1042

[3] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences*, vol. 11, no. 1, p. 237, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/1/237

[4] "Developing web-based support systems for predicting poor-performing students using educational data mining techniques," *International Journal of Advanced Computer Science and Applications*. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0110704

[5] "Find open datasets and machine learning projects — kaggle," 10.12.2024. [Online]. Available: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/data

[6] M. Asadozzaman, "Pathways to predicting student success," *Kaggle*, 01.10.2024. [Online]. Available: https://www.kaggle.com/code/asadozzaman/pathways-to-predicting-student-success#--10.-Model-Performance-Comparison---

[7] A. E. Ibrahem, "Student performance factors," *Kaggle*, 15.10.2024. [Online]. Available: https://www.kaggle.com/code/ahmedezzatibrahem/student-performance-factors

[8] S. Rafazi, "Student score with neural network," *Kaggle*, 18.10.2024. [Online]. Available: https://www.kaggle.com/code/alirafazi/student-score-with-neural-network#Step-4-%7C-Splitting-to-test-,-validation-set

[9] A. Gamal, "Studentperformance," *Kaggle*, 20.10.2024. [Online]. Available: https://www.kaggle.com/code/ahmedgmy/studentperformance/notebook

[10] W. Qi, C. Xu, and X. Xu, "Autogluon: A revolutionary framework for landslide hazard analysis," *Natural Hazards Research*, vol. 1, no. 3, pp. 103–108, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666592121000305

[11] AutoGluon, "autogluon.tabular.tabularpredictor," 27.11.2024. [Online]. Available: https://auto.gluon.ai/stable/api/autogluon.tabular.TabularPredictor.html

[12] ——, "autogluon.multimodal.multimodalpredictor," 27.11.2024. [Online]. Available: https://auto.gluon.ai/stable/api/autogluon.multimodal.MultiModalPredictor.html

[13] ——, "Autogluon multimodal - quick start," 27.11.2024. [Online]. Available: https://auto.gluon.ai/stable/tutorials/multimodal/multimodal_prediction/multimodal-quick-start.html

[14] B. Deepa and K. Ramesh, "Production level data pipeline environment for machine learning models," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, 2021, pp. 404–407. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9442035?casa_token=4B-sYMsYGsMAAAAA:tCaB3NbVVoOahfJ_ZR0LY9tnj2uko8UGkMNcffKv2K7OU5XqT0tg9WRZPhFUVm_w30AbC84atA

[15] W. He, X. Tang, W. Ji, L. Meng, J. Wei, D. Cao, C. Ma, Q. Li, and C. Lin, "An improved multi-island genetic algorithm and its utilization in the optimal design of a micropositioning stage," *Expert Systems with Applications*, vol. 257, p. 125029, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424018967

[16] "autogluon.core.models.ensemble.weighted_ensemble_model - autogluon 1.2.1 documentation," 11.12.2024. [Online]. Available: https://auto.gluon.ai/dev/_modules/autogluon/core/models/ensemble/weighted_ensemble_model.html

[17] AutoGluon, "How it works," 27.11.2024. [Online]. Available: https://auto.gluon.ai/stable/tutorials/tabular/how-it-works.html

[18] A. Verma, "Understanding common regression evaluation metrics: Mae, mse, rmse, r2, and adjusted r2," *Artificial Intelligence in Plain English*, 06.11.2023. [Online]. Available: https://ai.plainenglish.io/understanding-common-regression-evaluation-metrics-mae-mse-rmse-r2-and-adjusted-r2-6c5709e614c4