Explainable Artificial Intelligence Framework for Predicting Educational Inequality Using Random Forest and XGBoost Across Multiple Educational Levels Globally

Aulia Abikhair[1], Huang Guanghui[2,*]

[1]Teacher Education Study Program, School of Postgraduate Studies, Universitas Pendidikan Indonesia, Bandung, Indonesia

[2]Department of Educational Administration, Faculty of Education, Universitas Pendidikan Indonesia, Bandung, Indonesia

ABSTRACT

Educational inequality remains a major global challenge influenced by socioeconomic, regional, and developmental disparities across educational systems. However, accurately predicting educational inequality while maintaining model interpretability remains difficult due to the complex and nonlinear nature of educational data. This study proposes an explainable ensemble machine learning framework for predicting educational inequality across Primary, Lower Secondary, and Upper Secondary education levels using Random Forest and XGBoost models. The study employed regression evaluation metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), coefficient of determination (R^2), and five-fold cross-validation to evaluate model performance. The experimental results demonstrated that XGBoost achieved the best predictive performance with an R^2 value of 0.712 and a cross-validation mean score of 0.584, indicating strong predictive capability and stable generalization performance. To enhance model transparency, SHAP (SHapley Additive Explanations) analysis was applied to identify the most influential predictors contributing to educational inequality. The explainability results revealed that Education_Level_Upper Secondary, development-related variables, regional characteristics such as Sub-Saharan Africa, and temporal factors were the most significant contributors affecting prediction outcomes. In addition, the findings showed that educational inequality becomes increasingly distinguishable at higher educational levels, particularly within Upper Secondary education systems. The integration of Explainable Artificial Intelligence successfully improved model interpretability by providing transparent explanations regarding feature contributions and prediction behavior. Overall, this study demonstrates that explainable ensemble machine learning offers substantial potential for supporting evidence-based educational policy analysis, improving educational monitoring systems, and assisting policymakers in identifying critical factors contributing to educational inequality across diverse regional and educational contexts.

Keywords Explainable Artificial Intelligence, Educational Inequality Prediction, Ensemble Machine Learning, SHAP Explainability, Artificial Intelligence in Education.

Submitted 10 May 2026
Accepted 30 July 2026
Published 17 November 2026

*Corresponding author
Huang Guanghui,
huangguanghui@upi.edu

Additional Information and Declarations can be found on page 285

DOI: 10.63913/ail.v2i4.74
Copyright
2026 Abikhair and Guanghui

Distributed under
Creative Commons CC-BY 4.0

Introduction

Educational inequality remains one of the most significant global challenges affecting educational accessibility, participation, and quality across different countries and educational systems. Despite international efforts to improve

How to cite this article: A. Abikhair and H. Guanghui, "Explainable Artificial Intelligence Framework for Predicting Educational Inequality Using Random Forest and XGBoost Across Multiple Educational Levels Globally," *Artif. Intell. Learn.*, vol. 2, no. 4, pp. 268-287, 2026.

equal access to education, disparities in educational opportunities continue to exist due to socioeconomic conditions, regional development gaps, infrastructure limitations, and unequal resource distribution. These inequalities become increasingly visible at higher educational levels where financial barriers, school accessibility, and educational continuation challenges are more prominent. As educational systems continue to generate large-scale and multidimensional data, Artificial Intelligence and machine learning approaches have emerged as promising solutions for analyzing complex educational patterns and supporting data-driven educational policy development.

Recent developments in educational data mining and learning analytics have demonstrated that ensemble machine learning methods such as Random Forest and XGBoost are capable of producing strong predictive performance in various educational prediction tasks. These approaches have been widely applied for predicting student performance, dropout risk, online learning adaptability, academic achievement, and educational outcomes. Ensemble learning models are particularly effective because they can capture nonlinear relationships and interactions among multiple educational and socioeconomic variables. Furthermore, the integration of Explainable Artificial Intelligence techniques has gained increasing attention due to the growing need for transparency and interpretability in machine learning systems used within educational environments.

Although previous studies have shown promising results in educational prediction tasks, several important limitations remain. Most existing research primarily focuses on student-level predictions such as academic performance or dropout detection rather than broader educational inequality prediction at regional and educational system levels. In addition, many machine learning studies emphasize predictive accuracy while providing limited explanation regarding how specific features influence prediction outcomes. This lack of interpretability creates challenges for policymakers and educational stakeholders who require transparent and understandable insights before implementing evidence-based educational interventions. Furthermore, relatively few studies have investigated educational inequality prediction across multiple educational levels using explainable ensemble machine learning frameworks.

Another important issue is that educational inequality is strongly influenced by regional and developmental disparities that vary across countries and educational stages. Educational systems in less developed regions often experience limited infrastructure, insufficient educational funding, unequal access to learning opportunities, and higher dropout rates, particularly at secondary education levels. These conditions create highly complex and nonlinear educational inequality patterns that are difficult to analyze using conventional statistical methods. Consequently, there is a need for advanced predictive frameworks capable of handling multidimensional educational data while simultaneously providing transparent explanations regarding the most influential factors contributing to inequality patterns.

Therefore, this study proposes an explainable ensemble machine learning framework for predicting educational inequality across Primary, Lower Secondary, and Upper Secondary education levels using Random Forest and XGBoost models combined with SHAP Explainable Artificial Intelligence analysis. The proposed framework aims not only to achieve strong predictive performance but also to improve model transparency and interpretability by

identifying the most influential predictors contributing to educational inequality. In addition, this study evaluates predictive performance across different educational stages and investigates how regional, developmental, and temporal factors affect educational inequality outcomes. The findings of this study are expected to contribute to the advancement of Artificial Intelligence in Education by providing interpretable predictive analytics that can support educational monitoring systems, evidence-based policymaking, and targeted interventions for reducing educational inequality across diverse educational environments worldwide.

Literature Review and Related Works

Educational inequality has become a major concern in global education systems because unequal access to educational resources, infrastructure, and learning opportunities significantly affects educational outcomes and long-term socioeconomic development. Previous studies have shown that disparities in educational participation are strongly associated with regional development, household economic conditions, demographic factors, and educational accessibility [6], [7]. These inequalities become increasingly complex at higher educational levels, particularly in secondary education systems where financial limitations, dropout risks, and infrastructure gaps are more prominent [8]. As educational datasets continue to expand, Artificial Intelligence and machine learning methods have gained increasing attention for analyzing educational patterns and supporting evidence-based educational decision-making [9].

Machine learning techniques have been widely applied in educational data mining and learning analytics to predict student performance, academic achievement, dropout risk, online learning adaptability, and educational outcomes [10], [11]. Among these approaches, ensemble learning methods such as Random Forest, Gradient Boosting, and XGBoost have demonstrated strong predictive performance due to their ability to capture nonlinear relationships and multidimensional feature interactions [12], [13]. Previous studies have reported that ensemble machine learning models consistently outperform conventional statistical approaches in educational prediction tasks because they are more robust when handling heterogeneous educational and socioeconomic datasets [14]. In addition, ensemble methods have shown strong generalization performance through cross-validation and feature optimization techniques [15].

Despite the promising predictive capability of machine learning models, one of the major challenges in educational Artificial Intelligence is the black-box nature of many predictive systems. Traditional machine learning models often provide high predictive accuracy while offering limited transparency regarding how prediction decisions are generated [16]. This lack of interpretability creates difficulties for policymakers and educational stakeholders who require understandable explanations before implementing educational interventions. To address this issue, Explainable Artificial Intelligence has emerged as an important research area aimed at improving transparency, interpretability, and trustworthiness within machine learning systems [17]. Explainability techniques enable researchers to identify the contribution of individual features to prediction outcomes and improve understanding of model behavior in educational contexts [18].

Among various Explainable Artificial Intelligence techniques, SHAP (SHapley

Additive Explanations) has become one of the most widely adopted approaches for interpreting ensemble learning models. SHAP analysis provides both global and local interpretability by quantifying the contribution and directional influence of each feature on prediction outcomes [19]. Previous studies have successfully integrated SHAP with ensemble learning approaches to explain student performance prediction, dropout detection, learning achievement analysis, and educational adaptability prediction [20]. The integration of SHAP analysis allows researchers and policymakers to identify critical educational factors influencing prediction results while simultaneously improving transparency and reliability within educational Artificial Intelligence systems.

Although previous studies have demonstrated the effectiveness of machine learning and Explainable Artificial Intelligence approaches in educational prediction tasks, limited research has focused specifically on educational inequality prediction across multiple educational levels and regional development contexts. Most prior studies primarily investigate individual student-level outcomes rather than broader educational inequality patterns at regional or national levels. Furthermore, relatively few studies combine ensemble machine learning methods with SHAP explainability analysis to investigate educational inequality across Primary, Lower Secondary, and Upper Secondary education systems. Therefore, this study addresses these research gaps by proposing an explainable ensemble machine learning framework capable of predicting educational inequality while simultaneously providing interpretable insights regarding the most influential educational, regional, and developmental factors affecting prediction outcomes.

Methodology

Research Design

This study employed a quantitative predictive analytics approach integrating ensemble machine learning and Explainable Artificial Intelligence (XAI) techniques to predict educational inequality across multiple educational levels. The proposed framework was designed to analyze complex educational patterns by combining data preprocessing, feature engineering, machine learning modeling, model evaluation, and explainability analysis within a unified prediction pipeline. The primary objective of the methodology was not only to achieve strong predictive performance but also to improve transparency and interpretability regarding the factors contributing to educational inequality across different regional and developmental contexts. The overall research workflow consisted of several major stages, including dataset collection and integration, preprocessing, feature engineering, machine learning training, model evaluation, educational-level analysis, SHAP explainability analysis, and visualization. Figure 1 illustrates the overall research framework and methodological pipeline employed in this study.

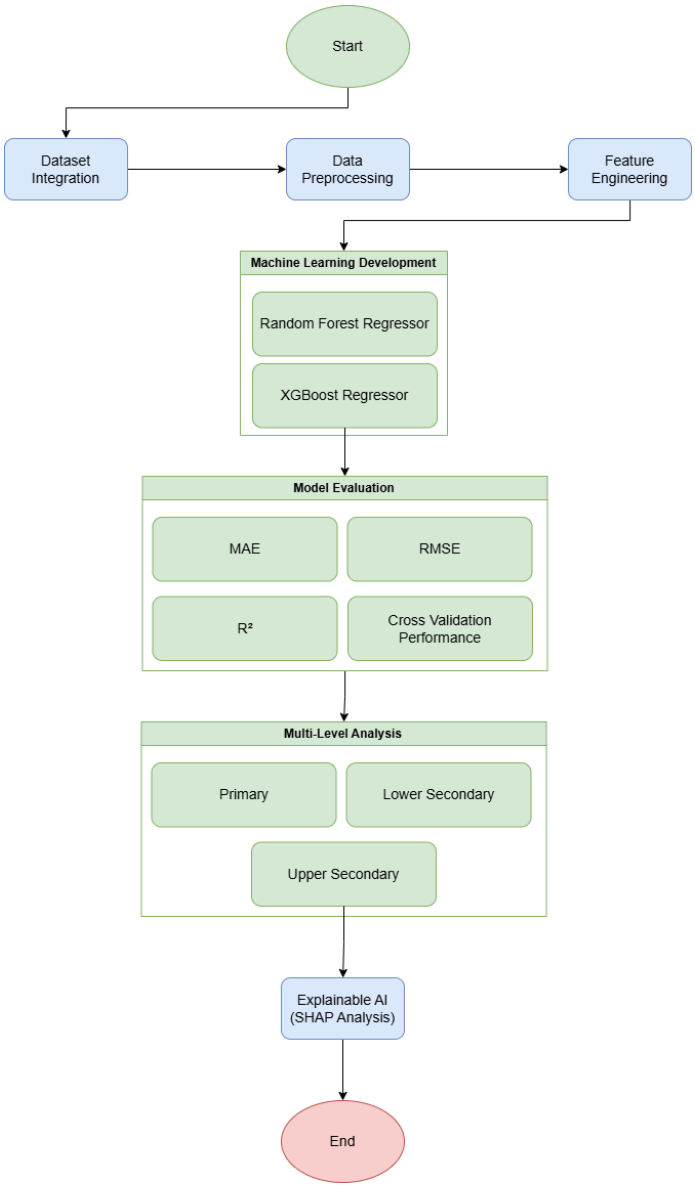


Figure 1 **Research Framework**

Dataset Collection and Integration

The dataset used in this study consisted of three separate educational datasets representing different educational stages, namely Primary Education, Lower Secondary Education, and Upper Secondary Education. These datasets were integrated into a unified dataset to enable comprehensive multi-level educational inequality analysis across diverse educational systems, countries, and regional development conditions. The integration process allowed the proposed machine learning framework to analyze educational inequality patterns across multiple educational stages within a single predictive environment. Each dataset contained educational, demographic, regional, and development-related indicators associated with educational participation and inequality conditions. The original variables included ISO3, Countries and areas, Region, Sub-region, Development Regions, Total, Female, Male, Rural_Residence, Urban_Residence, Poorest_Wealth quintile, Second_Wealth

quintile, Middle_Wealth quintile, Fourth_Wealth quintile, Richest_Wealth quintile, Data source, Time period, and Education_Level. To distinguish educational stages after integration, an additional categorical variable called Education_Level was assigned to each dataset with three categories: Primary, Lower Secondary, and Upper Secondary. This variable enabled the machine learning models to differentiate educational inequality characteristics across educational stages and supported multi-level educational performance analysis. After dataset integration, all column names were standardized by removing formatting inconsistencies and unnecessary whitespace characters to ensure preprocessing consistency, improve data quality, and facilitate efficient machine learning processing.

Data Preprocessing

Data preprocessing was conducted to improve data quality, reduce prediction bias, and ensure compatibility with machine learning algorithms. Several preprocessing procedures were systematically applied before model training.

To avoid direct target leakage and reduce model bias, several variables strongly associated with explicit educational inequality indicators were removed from the dataset during the preprocessing stage. These variables included Female, Male, Rural_Residence, Urban_Residence, Poorest_Wealth quintile, Second_Wealth quintile, Middle_Wealth quintile, Fourth_Wealth quintile, and Richest_Wealth quintile. These features were excluded because they directly represented demographic and socioeconomic inequality characteristics that could reveal target-related information and artificially inflate predictive performance. If retained, the machine learning models might rely excessively on these directly correlated indicators rather than learning broader educational inequality patterns derived from regional, developmental, and temporal relationships. By removing these variables, the proposed framework encouraged the models to identify more generalized and meaningful predictive patterns associated with educational inequality across different countries and educational levels. This preprocessing strategy also improved model robustness and reduced the risk of overfitting caused by highly correlated inequality indicators.

Several metadata variables were removed during preprocessing because they were not directly relevant to the predictive modeling process and could introduce unnecessary complexity into the machine learning framework. These variables included ISO3, Countries and areas, Data source, and Time period. The variables ISO3 and Countries and areas primarily served as country identifiers and did not provide meaningful predictive patterns for educational inequality analysis. Similarly, Data source was excluded because it represented administrative metadata rather than educational or socioeconomic characteristics related to prediction outcomes. Meanwhile, the Time period variable was initially removed from the original dataset structure and later transformed into a numerical temporal feature called Year during the feature engineering stage. This transformation enabled the machine learning models to capture longitudinal educational inequality trends and analyze temporal changes in educational conditions across different years more effectively.

Missing values were processed using different imputation strategies based on the characteristics of each feature type to improve data quality and maintain dataset consistency during model training. For numerical variables, median

imputation was applied because the median is more robust to outliers and extreme values compared to mean imputation. This approach helped preserve the overall numerical distribution while minimizing distortion caused by abnormal observations within the dataset. For categorical variables, most frequent imputation was employed by replacing missing values with the dominant category observed within each feature. This strategy preserved the most representative categorical patterns and reduced information loss during preprocessing. The use of separate imputation techniques for numerical and categorical variables enabled the machine learning models to process incomplete educational data more effectively while improving training stability and predictive reliability.

Numerical variables were standardized using StandardScaler to normalize feature distributions and improve training stability.

The standardization process was performed using the following equation:

$$z = \frac{x - \mu}{\sigma} x \tag{1}$$

$x$represents the original feature value,

μrepresents the feature mean,

σrepresents the standard deviation.

Categorical variables were transformed using One-Hot Encoding to convert categorical values into machine-readable numerical representations.

The preprocessing procedures were integrated into a unified pipeline using a ColumnTransformer architecture to ensure consistent preprocessing during both training and evaluation stages.

Feature Engineering

Feature engineering was conducted to improve the representation of regional development and temporal educational inequality characteristics.

The categorical variable Development Regions was transformed into a numerical ordinal variable called Dev_Index to represent the hierarchical development status of each country or region within the dataset. The transformation process applied ordinal mapping with the following values: Least Developed = 0, Less Developed = 1, and More Developed = 2. This feature engineering step was performed to convert categorical regional development information into a machine-readable numerical representation while preserving the ordinal relationship among development categories. By transforming development status into an indexed numerical variable, the machine learning models were able to capture hierarchical socioeconomic development patterns and analyze how different levels of national development influence educational inequality outcomes. This transformation also enabled the models to learn relationships between educational inequality and regional development conditions more effectively across multiple countries and educational contexts.

The Time period variable was converted into a numerical Year variable to capture longitudinal educational inequality dynamics and temporal educational changes across different years.

The target variable used in this study was Total, which represented the aggregated educational inequality indicator predicted by the machine learning

models. This variable reflected the overall educational inequality condition across different countries, regions, and educational stages within the dataset. The prediction of the Total variable enabled the machine learning framework to analyze how educational, regional, developmental, and temporal factors collectively influenced educational inequality patterns. By using Total as the target variable, the study aimed to capture broader inequality dynamics rather than focusing solely on individual demographic or socioeconomic indicators. This approach allowed the proposed machine learning models to learn complex relationships among multiple educational features and to provide a comprehensive representation of educational inequality conditions across Primary, Lower Secondary, and Upper Secondary education systems.

Machine Learning Models

This study employed two ensemble machine learning regression algorithms, namely Random Forest Regressor and XGBoost Regressor.

Random Forest is an ensemble learning algorithm based on multiple decision trees. The final prediction is generated by averaging predictions from multiple trees to reduce variance and improve robustness.

The prediction mechanism can be expressed as:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{2}$$

Brepresents the number of trees,

$f_b(x)$represents the prediction from each tree.

XGBoost is a gradient boosting ensemble algorithm that incrementally improves prediction performance by minimizing residual errors during training.

The objective function of XGBoost can be formulated as:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

$l(y_i, \hat{y}_i)$represents the loss function,

$\Omega(f_k)$represents regularization terms.

Training and Validation Strategy

The dataset was divided into training and testing subsets using an 80:20 ratio. The training dataset was used for model learning, while the testing dataset was used for evaluating predictive performance on unseen observations.

To improve robustness and generalization capability, five-fold cross-validation was applied during model evaluation. Cross-validation allowed the models to be trained and validated across multiple folds, thereby reducing the risk of overfitting.

The cross-validation score was calculated as:

$$CV = \frac{1}{K}\sum_{i=1}^{K} Score_i \tag{4}$$

Krepresents the number of folds,

$Score_i$represents the evaluation score for each fold.

Evaluation Metrics

The predictive performance of the machine learning models was evaluated using three regression metrics.

Mean Absolute Error (MAE)

MAE measures the average absolute difference between actual and predicted values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (5)$$

Root Mean Square Error (RMSE)

RMSE measures the square root of the average squared prediction error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (6)$$

Coefficient of Determination (R^2)

R^2 evaluates the proportion of variance explained by the machine learning model.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (7)$$

These evaluation metrics were selected because they provide complementary information regarding prediction accuracy, prediction error magnitude, and explanatory capability.

Educational-Level Performance Analysis

To investigate the robustness and consistency of the proposed prediction framework across different educational stages, separate prediction experiments were conducted for Primary Education, Lower Secondary Education, and Upper Secondary Education. Each educational level was independently filtered from the integrated dataset and processed using the same preprocessing, training, and evaluation procedures to ensure fair performance comparison across educational categories. Random Forest regression models were independently trained and evaluated for each educational level using an 80:20 train-test split configuration. The evaluation process employed MAE, RMSE, and R^2 metrics to measure prediction accuracy, prediction error magnitude, and explanatory capability for each educational stage. This educational-level analysis was designed to identify variations in educational inequality patterns and to examine whether predictive performance differed across educational stages. The analysis also enabled a deeper understanding of how educational inequality characteristics evolve from primary to upper secondary education systems. By evaluating prediction performance separately for each educational level, the study was able to determine which educational stage exhibited more distinguishable inequality patterns and stronger machine learning predictability.

Explainable Artificial Intelligence Analysis

To improve model transparency and interpretability, SHAP (SHapley Additive Explanations) analysis was applied to the Random Forest model. SHAP analysis was selected because it provides both global and local interpretability

regarding feature contributions to prediction outcomes.

The SHAP value calculation can be expressed as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{i\}) - f(S)] \quad (8)$$

$\phi_i$represents the SHAP value of feature i,

Srepresents feature subsets,

Frepresents the complete feature set.

The explainability analysis included:

SHAP Summary Plot

SHAP Feature Importance Bar Plot

SHAP Dependence Plot

This analysis enabled the identification of the most influential educational, regional, developmental, and temporal variables affecting educational inequality predictions.

Results

Model Performance Evaluation

As presented in Table 1, this study evaluated the predictive performance of Random Forest and XGBoost models in predicting educational inequality using several regression evaluation metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), coefficient of determination (R^2), and five-fold cross-validation. The experimental results revealed that XGBoost achieved the highest predictive performance with an R^2 value of 0.712, indicating that the model was capable of explaining approximately 71.2% of the variance in educational inequality data. In addition, the model obtained a cross-validation mean score of 0.584, demonstrating relatively stable performance across different validation folds. The RMSE value of 10.842 further indicates that the prediction error remained within an acceptable range for educational and socioeconomic prediction tasks. These findings suggest that XGBoost successfully captured the complex relationships among educational level, regional characteristics, development status, and temporal variables within the dataset.

As further indicated by the comparative results in Table 1, although the Random Forest model achieved a slightly lower MAE value compared to XGBoost, its overall predictive capability was lower due to reduced R^2 performance and higher RMSE values. The superior performance of XGBoost may be attributed to its gradient boosting mechanism, which iteratively improves prediction accuracy by minimizing residual errors during training. Consequently, XGBoost was selected as the primary prediction model for educational inequality forecasting. However, Random Forest was employed for Explainable Artificial Intelligence analysis because of its more stable feature importance structure and compatibility with SHAP interpretation techniques. This combination enabled the study not only to achieve strong predictive accuracy but also to provide transparent and interpretable insights into the key factors influencing educational inequality across multiple educational levels.

Table 1 Performance Comparison of Machine Learning Models

Model	MAE	RMSE	R^2	CV R^2 Mean	CV R^2 Std
Random Forest	7.293	11.065	0.700	0.526	0.110
XGBoost	7.520	10.842	0.712	0.584	0.136

Actual and Predicted Value Analysis

Figure 2 illustrates the relationship between actual and predicted educational inequality values generated by the XGBoost model. The scatter plot demonstrates that most prediction points were concentrated near the diagonal reference line, indicating strong agreement between actual and predicted values. This pattern suggests that the model effectively learned the underlying relationships within the dataset and successfully captured the overall distribution of educational inequality across different countries and educational levels. The relatively close alignment of the prediction points with the reference line also indicates that the model achieved consistent predictive accuracy for the majority of observations. However, several deviations were observed at higher actual values, particularly for educational inequality scores above 50, where prediction errors became more noticeable. This finding indicates that extreme inequality conditions remain more complex and difficult to model accurately due to the presence of nonlinear socioeconomic and regional factors. Despite these deviations, the overall visualization confirms that the proposed ensemble learning approach produced robust and reliable prediction performance with strong predictive consistency across most educational inequality observations.

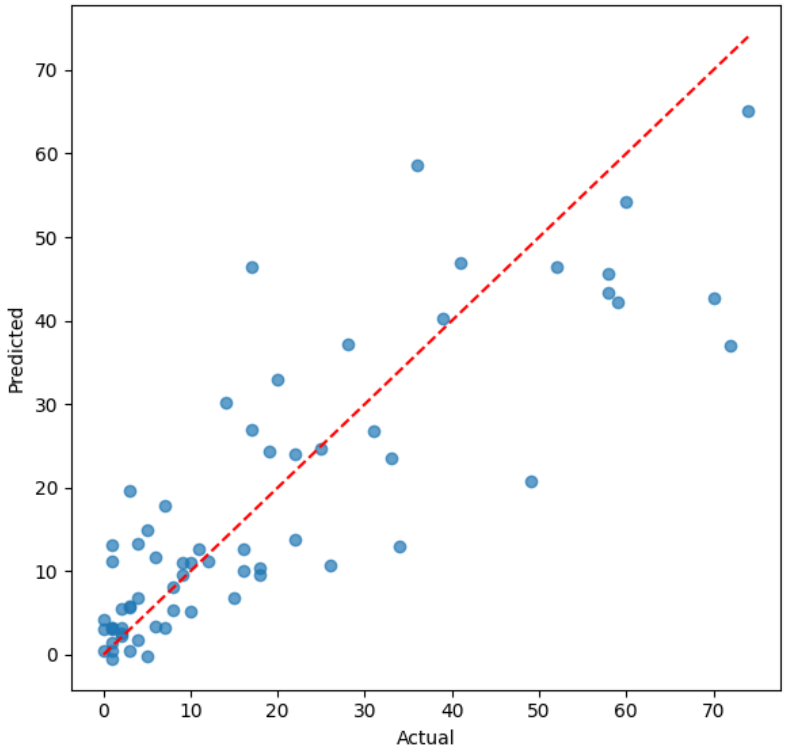


Figure 2 Actual vs Predicted Values Using XGBoost

Prediction Error Distribution

Figure 3 presents the prediction error distribution generated by the proposed

XGBoost model for educational inequality prediction. The histogram shows that the majority of prediction errors were concentrated around zero, indicating that the model produced predictions that were generally close to the actual values. The distribution also followed a relatively normal pattern, suggesting that the prediction errors were balanced and randomly distributed rather than systematically biased toward overprediction or underprediction. Most error values were located within a moderate range, which demonstrates that the model maintained stable prediction behavior across different observations in the dataset. In addition, the absence of extreme skewness in the distribution indicates that the model was capable of handling diverse educational inequality patterns without generating excessive prediction instability. Although several larger errors were still observed, their frequency remained relatively low compared to the overall distribution. These findings confirm that the proposed ensemble learning approach achieved robust predictive stability and reliable generalization performance across multiple educational contexts and educational levels.

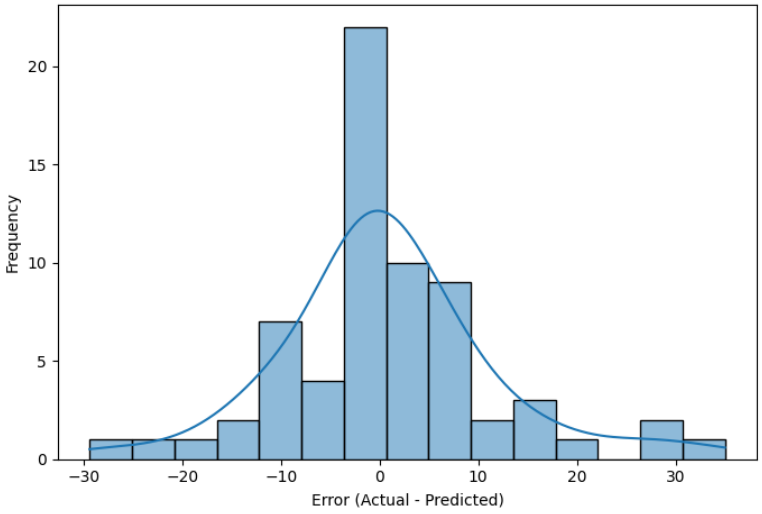


Figure 3 **Prediction Error Distribution**

Performance Across Educational Levels

As presented in table 2, to evaluate model robustness across different educational stages, prediction performance was analyzed separately for Primary, Lower Secondary, and Upper Secondary education levels using MAE, RMSE, and R^2 metrics. The experimental results revealed that the model achieved the highest predictive performance at the Upper Secondary level with an R^2 value of 0.451, followed by Lower Secondary with an R^2 value of 0.388, while the Primary level produced the lowest performance with an R^2 value of 0.299. These findings indicate that educational inequality patterns become increasingly distinguishable and predictable at higher educational levels. The stronger predictive performance observed in Upper Secondary education may reflect the presence of more pronounced disparities related to educational access, socioeconomic conditions, school infrastructure, dropout rates, and regional development gaps. In many developing regions, participation in upper secondary education is often more heavily influenced by household income, educational policy, and geographic accessibility, making inequality patterns more visible within the data. In contrast, primary education systems are

generally more universal and publicly supported, which may reduce observable inequality variation and limit the model's ability to distinguish complex predictive patterns. Overall, the results presented in Table 2 demonstrate that educational inequality exhibits varying characteristics across educational stages and becomes increasingly complex at higher levels of education.

Table 2 Prediction Performance Across Educational Levels

Education Level	MAE	RMSE	R^2
Primary	4.544	8.460	0.299
Lower Secondary	9.382	13.328	0.388
Upper Secondary	11.342	13.355	0.451

Feature Importance Analysis

As illustrated in figure 4, feature importance analysis was conducted using the Random Forest model to identify the most influential predictors contributing to educational inequality prediction. The results revealed that Education_Level_Upper Secondary was the most dominant feature, followed by Year, Region_SSA, Dev_Index, and Development Regions_Least Developed. The high importance of Education_Level_Upper Secondary indicates that educational inequality becomes increasingly pronounced at higher educational stages, where disparities in access, affordability, infrastructure, and educational continuation rates are generally more visible. This finding suggests that upper secondary education systems are more sensitive to socioeconomic and regional inequalities compared to lower educational levels. In addition, regional variables such as Region_SSA demonstrated strong predictive influence, highlighting the substantial contribution of Sub-Saharan African countries to global educational inequality patterns. Development-related variables, including Dev_Index and Development Regions_Least Developed, also showed considerable importance, confirming that national development status strongly affects educational participation and equality outcomes. Countries with lower development levels are more likely to experience limited educational infrastructure, unequal resource distribution, and restricted educational opportunities, which directly contribute to higher inequality levels. Furthermore, the importance of the Year variable indicates that educational inequality patterns evolve over time and may be influenced by changing socioeconomic conditions, educational policies, demographic transitions, and global events affecting education systems across multiple regions.

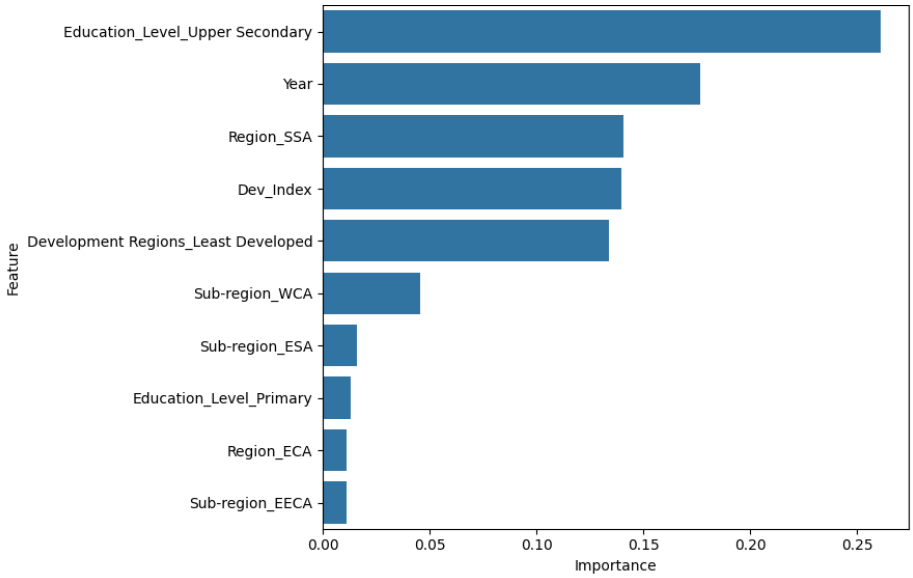


Figure 4 Top 10 Important Features Using Random Forest

SHAP Explainability Analysis

To improve model transparency and interpretability, SHAP (SHapley Additive Explanations) analysis was applied to the Random Forest model. Figure 5 presents the SHAP summary plot, which illustrates the overall contribution and directional impact of each feature on educational inequality prediction outcomes. The SHAP analysis revealed that the most influential features affecting prediction performance were Education_Level_Upper Secondary, Dev_Index, Region_SSA, Development Regions_Least Developed, and Year. Among these variables, Education_Level_Upper Secondary produced the strongest positive SHAP impact on the model output. Higher feature values consistently increased educational inequality prediction scores, indicating that inequality patterns become substantially more visible and pronounced at upper secondary education levels. This finding aligns with the feature importance analysis and suggests that disparities related to educational access, affordability, dropout risk, and educational continuation become increasingly significant in higher educational stages.

The SHAP summary plot also demonstrated that regional and development-related variables contributed substantially to model predictions. In particular, Region_SSA generated strong positive SHAP values, indicating that countries located in Sub-Saharan Africa significantly contributed to higher predicted educational inequality levels. This result reflects persistent educational challenges within developing regions, including unequal educational infrastructure, limited educational resources, socioeconomic instability, and disparities in educational accessibility. Similarly, Dev_Index and Development Regions_Least Developed exhibited considerable SHAP influence, confirming that national development status plays a critical role in shaping educational inequality patterns. Countries categorized as less developed tend to experience higher educational disparities due to economic limitations, insufficient public educational investment, and unequal distribution of learning opportunities. In addition, the Year variable demonstrated meaningful SHAP contributions, suggesting that educational inequality evolves over time and may be influenced

by policy changes, demographic transitions, technological development, and global educational conditions.



Figure 5 SHAP Summary Plot

Figure 6 presents the SHAP feature importance bar plot, which further confirms the consistency of global feature importance rankings identified in the summary plot. Education_Level_Upper Secondary remained the most influential predictor, followed by development index and regional variables. The consistency between the SHAP summary plot and SHAP bar plot strengthens the reliability of the explainability analysis and demonstrates that the proposed machine learning framework successfully identified meaningful predictors of educational inequality. Overall, the SHAP analysis significantly enhanced model interpretability by providing transparent explanations regarding feature contributions and prediction behavior. These findings highlight the potential of explainable artificial intelligence for supporting evidence-based educational policy analysis, improving trust in machine learning predictions, and assisting policymakers in identifying the most critical factors contributing to educational inequality across multiple educational levels and regional contexts.

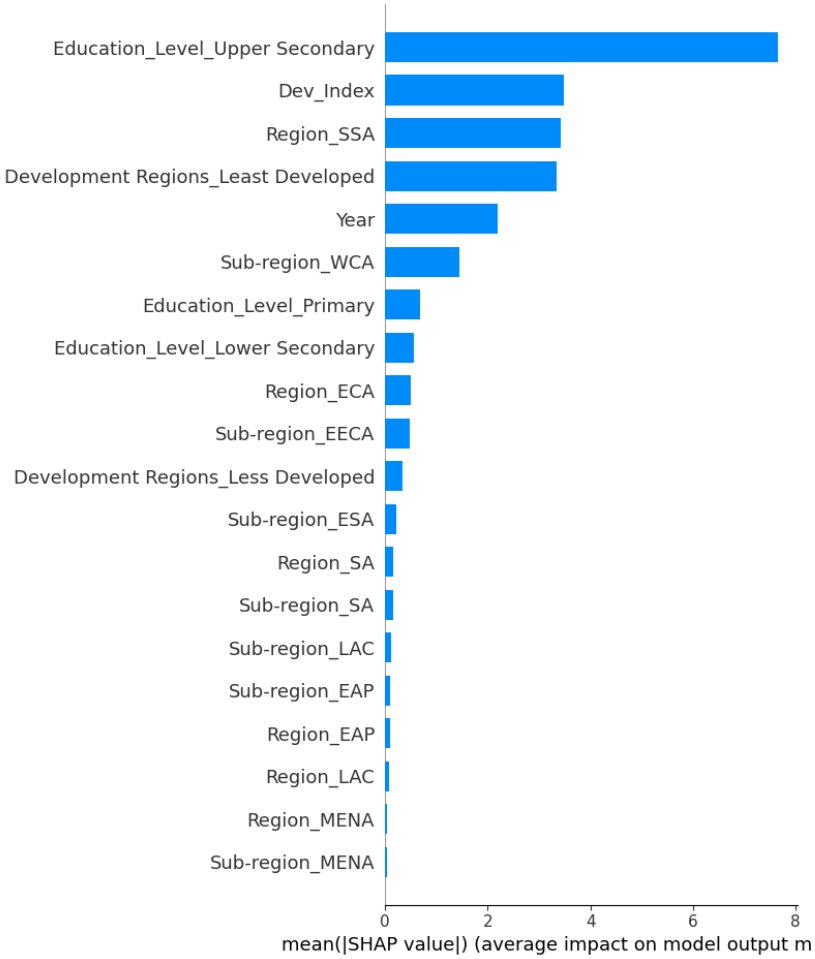


Figure 6 SHAP Feature Importance Bar Plot

Discussion

The findings of this study demonstrate that ensemble machine learning models are highly effective for predicting educational inequality across multiple educational levels and regional contexts. Among the evaluated models, XGBoost achieved the best overall predictive performance with an R^2 value of 0.712, indicating that the model was capable of explaining a substantial proportion of variance within the educational inequality dataset. This result suggests that educational inequality patterns are strongly associated with complex interactions among educational level, regional characteristics, development status, and temporal conditions. The superior performance of XGBoost may be attributed to its gradient boosting mechanism, which incrementally improves prediction accuracy by reducing residual prediction errors during training. This finding is consistent with recent educational data mining studies showing that ensemble models, especially XGBoost and Random Forest, are effective for predicting educational outcomes and identifying complex nonlinear relationships in student and institutional data [21], [22]. In addition, the relatively stable cross-validation performance indicates that the model maintained strong generalization capability despite the diversity and heterogeneity of the dataset. These findings confirm that ensemble machine learning approaches are suitable for handling educational and socioeconomic

prediction problems characterized by nonlinear relationships and multidimensional feature interactions. The results also demonstrate the growing potential of artificial intelligence methods within educational data mining and learning analytics for supporting predictive educational policy analysis [23].

The analysis across educational stages revealed that predictive performance became stronger at higher educational levels, particularly within Upper Secondary education. This finding indicates that educational inequality becomes increasingly distinguishable as students progress to more advanced educational stages. Upper secondary education systems are often more strongly affected by socioeconomic disparities, educational accessibility, household income limitations, infrastructure availability, and school continuation challenges. As a result, inequality patterns become more visible and easier for machine learning models to identify. In contrast, primary education systems are generally more universal and publicly supported, resulting in lower observable variability in inequality patterns. Furthermore, the feature importance and SHAP analyses consistently identified Education_Level_Upper Secondary, Region_SSA, and development-related variables as the most influential predictors affecting educational inequality. These findings suggest that less developed regions, particularly countries located in Sub-Saharan Africa, continue to experience substantial educational disparities associated with limited educational resources, uneven infrastructure distribution, and socioeconomic instability. Previous explainable machine learning research in education also shows that socioeconomic, institutional, demographic, and infrastructure-related variables are frequently important predictors of student performance, dropout, and educational risk [22], [24]. The integration of SHAP explainability analysis also represents an important contribution of this study because it enhances model transparency and allows policymakers to understand how specific features contribute to prediction outcomes. Unlike conventional black-box artificial intelligence approaches, the proposed explainable framework provides interpretable insights that can support evidence-based educational decision-making, improve trust in machine learning systems, and assist governments and educational institutions in identifying critical factors contributing to educational inequality across different educational contexts [21], [25].

Conclusion

This study proposed an explainable ensemble machine learning framework for predicting educational inequality across Primary, Lower Secondary, and Upper Secondary education levels using Random Forest, XGBoost, and SHAP Explainable Artificial Intelligence techniques. The experimental results demonstrated that XGBoost achieved the best predictive performance with an R^2 value of 0.712, indicating that the proposed model successfully captured complex educational inequality patterns across multiple regional and developmental contexts. The findings further revealed that educational inequality becomes increasingly distinguishable at higher educational stages, particularly within Upper Secondary education, where socioeconomic disparities, educational accessibility limitations, and infrastructure gaps tend to be more pronounced. Feature importance and SHAP analyses consistently identified Education_Level_Upper Secondary, Region_SSA, development-related indicators, and temporal variables as the most influential predictors affecting educational inequality outcomes. These results confirm that educational inequality is strongly associated with regional disparities and

national development conditions, especially in less developed regions such as Sub-Saharan Africa. In addition, the integration of SHAP analysis significantly improved model transparency by providing interpretable explanations regarding feature contributions and prediction behavior, thereby addressing the black-box limitations commonly associated with artificial intelligence models. Overall, this study demonstrates that explainable ensemble machine learning offers substantial potential for supporting evidence-based educational policy analysis, improving educational monitoring systems, and assisting policymakers in identifying critical factors contributing to educational inequality across diverse educational environments worldwide.

Declarations

Author Contributions

Conceptualization: A.A., H.G.; Methodology: A.A.; Software: H.G.; Validation: A.A., H.G.; Formal Analysis: A.A., H.G.; Investigation: H.G.; Resources: A.A.; Data Curation: A.A.; Writing Original Draft Preparation: A.A., H.G.; Writing Review and Editing: A.A., H.G.; Visualization: H.G.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, andor publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Global Education Monitoring Report Team, *Global Education Monitoring Report 2023: Technology in Education: A Tool on Whose Terms?* Paris, France: UNESCO, 2023, doi: 10.54676/UZQV8501.

[2] S. Marginson, “Limitations of human capital theory,” *Studies in Higher Education*, vol. 44, no. 2, pp. 287–301, 2019, doi: 10.1080/03075079.2017.1359823.

[3] S. Fujihara, “Identifying the role of high school in educational inequality: A causal mediation approach,” *Social Science Research*, vol. 124, no. November, Art. no. 103077, pp. 1-18, Nov. 2024, doi: 10.1016/j.ssresearch.2024.103077.

[4] R. S. Baker and P. S. Inventado, “Educational data mining and learning analytics,” in *Learning Analytics*, J. A. Larusson and B. White, Eds. New York, NY, USA:

Springer, vol. 2014, no. January, pp. 61–75, 2014, doi: 10.1007/978-1-4614-3305-7_4.

[5] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, “The current landscape of learning analytics in higher education,” *Computers in Human Behavior*, vol. 89, no. December, pp. 98–110, 2018, doi: 10.1016/j.chb.2018.07.027.

[6] S. Lamb, E. Markussen, R. Teese, J. Polesel, and N. Sandberg, Eds., *School Dropout and Completion: International Comparative Studies in Theory and Policy*. Dordrecht, The Netherlands: Springer, 2011, doi: 10.1007/978-90-481-9763-7.

[7] H. G. van de Werfhorst and J. J. B. Mijs, “Achievement inequality and the institutional structure of educational systems: A comparative perspective,” *Annual Review of Sociology*, vol. 36, no. August, pp. 407–428, 2010, doi: 10.1146/annurev.soc.012809.102538.

[8] S. Goldrick-Rab, “Challenges and opportunities for improving community college student success,” *Review of Educational Research*, vol. 80, no. 3, pp. 437–469, 2010, doi: 10.3102/0034654310370163.

[9] A. Peña-Ayala, “Educational data mining: A survey and a data mining-based analysis of recent works,” *Expert Systems with Applications*, vol. 41, no. 4, pt. 1, pp. 1432–1462, 2014, doi: 10.1016/j.eswa.2013.08.042.

[10] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[11] J. Xu, K. H. Moon, and M. van der Schaar, “A machine learning approach for tracking and predicting student performance in degree programs,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 742–753, 2017, doi: 10.1109/JSTSP.2017.2692560.

[12] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. October, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[13] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, vol. 2016, no. August, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[14] A. M. Shahiri, W. Husain, and N. A. Rashid, “A review on predicting student’s performance using data mining techniques,” *Procedia Computer Science*, vol. 72, no. 2015, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[15] F. Pedregosa *et al*., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html

[16] Z. C. Lipton, “The mythos of model interpretability,” *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018, doi: 10.1145/3233231.

[17] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” *IEEE Access*, vol. 6, no. September, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[18] D. Gunning and D. W. Aha, “DARPA’s explainable artificial intelligence (XAI) program,” *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019, doi: 10.1609/aimag.v40i2.2850.

[19] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *arXiv* preprint arXiv:1705.07874, vol. 2017, no. May, pp. 1-10, 2017, doi: 10.48550/arXiv.1705.07874.

[20] H. Sahlaoui, E. A. A. Alaoui, A. Nayyar, S. Agoujil and M. M. Jaber, "Predicting and Interpreting Student Performance Using Ensemble Models and Shapley Additive Explanations," in *IEEE Access*, vol. 9, pp. 152688-152703, 2021, doi: 10.1109/ACCESS.2021.3124270.

[21] T. Tong and Z. Li, “Predicting learning achievement using ensemble learning with result explanation,” *PLOS ONE*, vol. 20, no. 1, Art. no. e0312124, 2025, doi: 10.1371/journal.pone.0312124.

[22] R. Guevara-Reyes, I. Ortiz-Garcés, R. O. Andrade, F. Cox-Riquetti, and W. Villegas-Ch., “Machine learning models for academic performance prediction: Interpretability and application in educational decision-making,” *Frontiers in Education*, vol. 10, no. August, Art. no. 1632315, 2025, doi: 10.3389/feduc.2025.1632315.

[23] N. R. Raji, R. M. S. Kumar, and C. L. Biji, “Explainable machine learning prediction for the academic performance of deaf scholars,” *IEEE Access*, vol. 12, no. February, pp. 23595–23612, 2024, doi: 10.1109/ACCESS.2024.3363634.

[24] M. Quimiz-Moreira, R. Delgadillo, J. Parraga-Alava, N. Maculan, and D. Mauricio, “Factors, prediction, explainability, and simulating university dropout through machine learning: A systematic review, 2012–2024,” *Computation*, vol. 13, no. 8, Art. no. 198, 2025, doi: 10.3390/computation13080198.

[25] M. Arifin, F. Nugraha, and D. L. Fithri, “Student performance classification using academic, socioeconomic, and digital behavior features: A comparative study,” *Journal of Information Systems and Informatics*, vol. 8, no. 1, Feb. 2026, doi: 10.63158/journalisi.v8i1.1460.