JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Regression-Based Machine Learning Approach for Student Performance Prediction with Cross-Validation Stability

[1]S.Suvetha, [2] Dr.C.Immaculate Mary
[1]Research Scholar, [2]Associate Professor and Head,
[1] PG and Research Department of Computer Science,
[1]Sri Sarada College for Women(A), Salem, India

Abstract: In educational data mining (EDM) student performance prediction had evolved as an important research domain. The primary aim is to predict the student outcome and to analyse the student at risk which helps the institutions to improve the education quality and makes better decisions. This study proposes a structured analysis which focuses on predicting student marks using different regression-based machine learning models. The dataset, sourced from the UCI Machine Learning Repository it incorporates demographic, socioeconomic, and academic factors as predictive features. Multiple regression algorithms, such as Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor were implemented and evaluated using cross-validation technique and residual analysis. Each model performance was analyzed through various performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The results indicate that Lasso Regression achieved the highest predictive accuracy ($R^2 = 0.874$, RMSE = 1.18) which outperforms other models, while ensemble methods such as Random Forest and Gradient Boosting also delivered competitive results. In contrast, Decision Tree exhibited lower performance because of high error rates.

Index Terms **- Educational data mining, Machine learning, Student performance prediction, Regression models, Residual analysis, Model comparison.**

I. INTRODUCTION

The educational data ming (EDM) is one of the prominent research fields which completely focuses on the applying the data to the machine learning and statistical techniques to analyse and gather the insights in the educational domain. The primary aim of the educational data mining is to reveal the hidden patterns, find the relationships which in turns helps to make valuable decision making and enhance teaching. As the digital platform is growing the educational data mining is emerged as a powerful tool for making evidence-based improvements. The primary aim of the educational data mining is to make student performance prediction. The models are developed and deployed to forecast the unseen insights and academic outcomes such as grades, retention rate of students and their likelihood. By investigating the academic background and behavioral background of the student who are at risk can be identified easily. The model prediction not only helps to find the student academic prediction also aids the educators and institutions to understand the key driving factors like enabling the personalize learning plans.

The work flow is designed on three phrases. First data preprocessing, second model training with evaluations, third error analysis. The first step data preprocessing involves data cleaning, label encoding, handling categorical values, dealing with missing and duplicate values and preparing the dataset for complete training and testing. In the second stage of model training various machine learning models were implemented. In this work various regression models with regularization and liner models and ensemble

models were used. The models such as lasso regression, liner regression, ridge regression, decision tree, gradient boosting, random forest and XG boost algorithm were implemented. Each and every model is trained with the train split of 70% of data and results are noted. The efficiency of model is analyzed with various metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). in the final step of error analysis cross fold validation and residual analysis is carried out to make sure the predictions are accurate.

II. LITERATURE REVIEW

In [1], the author used various models such as XGBoost against Random Forest (RF), Lasso Regression, Elastic Net, Support Vector Machine, and Decision Tree. The result showed that XGBoost consistently outperforms other models with high accuracy and high improvements in the R^2 score ranging from 6.3% to 12.1% compared to other models. The work also includes feature engineering.

In [2], the author has used a dataset student performance prediction is taken from UCI repository. Classification techniques were implemented with metrics precision, recall and f1score. Author proposed work with deep learning neural network with 87.4% accuracy and Random Forest with 85.6% accuracy is highlighted.

In [3], comparative analysis among all the linear and regularization regression models were implemented with the evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2. ensemble methods and support vector regression seems to have similar level of performance.

In [4], the regression models like Linear **Regression, Decision Trees, Random Forests, Gradient Boosting, XGBoost, and Neural Networks** was deployed using student academic data. Results emphasized that Support Vector Regression and Linear Regression provided reliable predictive accuracy, representing the robustness of regression-based methods in educational contexts.

In [5], the dataset has been taken from UCI repository and the **Decision Tree model** was applied for predicting student dropout and academic performance. The model achieved 71% accuracy in classifying students as dropout, enrolled, or graduate. Course performance and tuition payment status were identified as Key predictors demonstrating the interpretability of tree-based methods.

III. METHODOLOGY

The methods implemented in this research work is designed with a proper structure that begins with dataset preprocessing. Then the process of the training and evaluation of multiple machine learning models was tested and results are evaluated. Finally with the residual plot, histogram analysis and cross validation techniques error were analysed. The work was implemented in Python-jupyter notebook using libraries such as pandas for data frames and data preprocessing, scikit-learn for model implementations, matplotlib for data visualizations, and XGBoost. Each step in the methodology ensures that the resulting predictions are both accurate and reliable, while minimizing biases and inconsistencies that often plague educational datasets.

3.1 Dataset Description

The dataset used in this study is taken from UCI repository. It contains academic, demographic and behavioural details of the students. The academic data such as their grade 1, grade 2 and grade 3 marks. Behavioural data such as study time, interest, activities, school up, travel time, guardian details, father mother education and occupation. The demographic details like age, higher, internet are included. Some features have numerical features and some features have categorical values. In total, there are 649 student records with 33 attributes.

3.2 Data Preprocessing

Data preprocessing on one of the curious step in data analytics because if it is not handled properly it might lead to inconsistent and invalid results to make our prediction more accurate the preprocessing is essential. Raw data will have many null, missing values and duplicate values. since the dataset is taken from UCI repository student performance prediction does not have many null values. Label encoding is also done with the help of standard scalar and one hot encoding. Later the dataset is divided for testing (y) and training(x). Finally, the dataset was splitted in 70:30 ratio for testing and training. 70% of data would go for model training and 30% will go for testing.

3.3 Model Selection

Linear Regression: By simulating a straight-line relationship between input features and the target variable, linear regression provides a baseline for continuous outcome prediction. To estimate coefficients, the sum of squared residuals is minimized. It is easy to understand, but it makes the assumption that the predictors are independent and linear. When relationships are extremely non-linear or there is multicollinearity in the data, its efficacy may be restricted.

Ridge Regression: By including L2 regularization, which penalizes large coefficients, Ridge Regression builds upon Linear Regression. This reduces variance and helps control overfitting, especially when predictors are highly correlated. All features remain in the model, but their impact is reduced proportionately. When there are a lot of correlated variables in the dataset, it is especially helpful.

Decision Tree Regression: Decision Tree Regression is useful for identifying non-linear patterns because it divides data into regions according to feature thresholds. It can easily handle both categorical and numerical features. Single trees, however, frequently overfit, resulting in erratic predictions with a high variance. They are nevertheless interpretable and serve as the foundation for numerous ensemble approaches.

The Random Forest Regressor: Several Decision Trees constructed using bootstrapped samples and random feature subsets are combined by Random Forest. This ensemble preserves accuracy while lowering variance and overfitting. It can successfully capture intricate feature interactions and is resilient to noisy data. The model's excellent generalization performance across a variety of datasets has earned it widespread recognition.

Gradient Boosting : Gradient Boosting builds trees one after the other, fixing the mistakes of the previous tree. High predictive accuracy and robustness are produced by this iterative learning process. To prevent overfitting, it necessitates meticulous adjustment of parameters such as learning rate and depth. For structured data, gradient boosting is effective and frequently outperforms more straightforward models.

XGBoost Regressor: A refined version of gradient boosting, XGBoost adds sophisticated regularization and computational efficiency. It is well-liked in competitive machine learning because of its accuracy, scalability, and speed design. It efficiently manages big datasets thanks to parallelization and integrated cross-validation. It is a cutting-edge boosting framework because of its versatility and excellent performance.

Lasso Regression: The regression technique that enhances the model's performance by incorporating L1 regularization is called least absolute shrinkage and selection operator (lasso regression), which places a limit on the absolute magnitude of the regression coefficients. By reducing a few feature coefficients to precisely zero, this restriction helps to simplify the model.

$$\min_{\beta} \sum_{k=1}^{m} (y_j - \hat{y}_j)^2 + \lambda \sum_{d=1}^{q} |\beta_j|$$

where:

- y_j represents the actual values, and $\hat{y}_j$ represents the predicted values,
- β_j are the regression coefficients,
- λ is the regularization parameter that controls the amount of shrinkage.

3.4 Evaluation Metrics

Multiple error and accuracy metrics were used to assess the regression models' performance. These metrics measure each model's efficiency as well as the difference between expected and actual student grades.

Mean Squared Error (MSE):

The average squared difference between expected and actual values is measured by the Mean Squared Error, or MSE. Larger deviations are penalized more severely because errors are squared, which makes it susceptible to outliers.

Root Mean Squared Error (RMSE):

The square root of MSE, expressed in the same units as the target variable, is known as the root mean squared error, or RMSE. It is frequently used to compare model accuracy and offers a more comprehensible indicator of prediction error magnitude.

Mean Absolute Error (MAE):

The average absolute difference between expected and actual values is calculated by the Mean Absolute Error, or MAE. It provides a reliable indicator of model performance since it handles all errors equally and is less inclined to outliers than RMSE.

Coefficient of Determination (R2):

R2 is the percentage of the target variable's variance that the model can account for. Stronger predictive fit is indicated by values near 1, while poor explanatory power is indicated by values close to 0.

Cross- Validation:

K-fold cross-validation was used to ensure robustness. In particular, tests were carried out using various fold values (k = 2, 3, 4, 5, 6), and the mean RMSE and R2 values for each fold were reported. This method mitigates bias, lessens reliance on a single train-test split, and offers a more accurate assessment of model generalization.

IV. RESULTS

The outcomes of various regression models to the student dataset for predicting their performance is demonstrated below. The model performance with the MSE, MAE and R^2 metrics for each model is tabulated. The residual analysis, residual histogram analysis and finally cross validation stability across all models are analyzed.

4.1 Model Performance

The results of different regression models are tabulated in below table1. From the results it is clear that lasso regression has lower error value of 1.394 and highest R^2 value of 0.8738. whereas Decision tree has very highest error rate of MSE of 4.892 and very lowest R^2 value of 0.5575 it confirms the tendency of overfitting and also lacks in robustness on unforeseen data. On the other stand ensemble methods such as Random Forest have achieved a MSE = 1.533 and R^2 = 0.861, showing stable and reliable predictions across features. Gradient Boosting algorithm shows an comparative results slightly better than Random Forest model. XG boost lagged behind while predicted across all other regression models. Finally, the results showcase that the regularized linear models lasso regression and ensemble model Random Forest are more efficient in detecting the unseen patterns in student performance prediction.

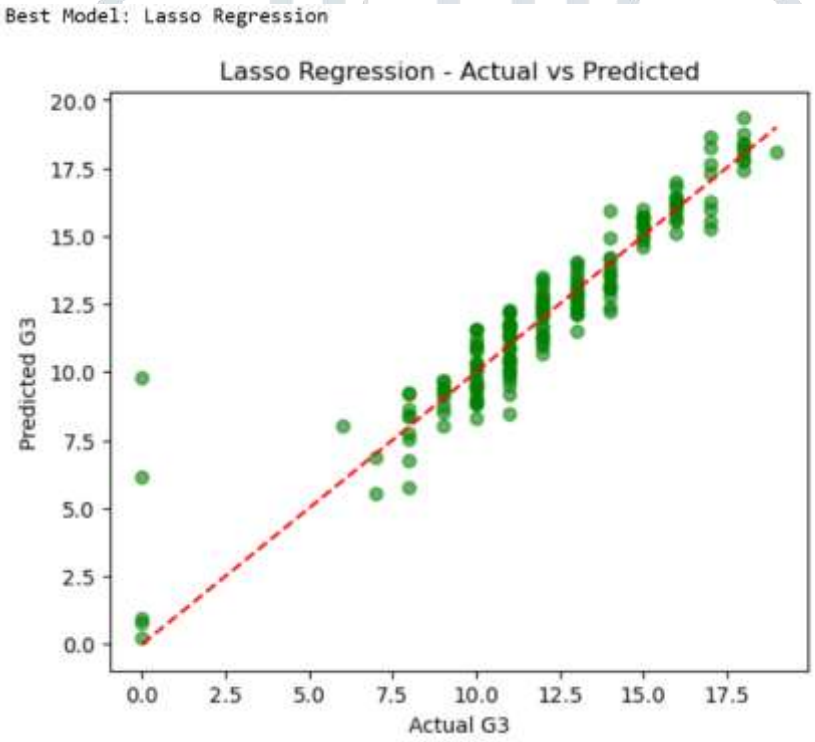
Table 1: performance of models

Model	MSE	RMSE	MAE	R^2
Linear Regression	1.438610	1.199421	0.784661	0.869905
Ridge Regression	1.439156	1.199648	0.784588	0.869856

Lasso Regression	**1.394648**	1.180952	0.771556	**0.873880**
Decision Tree	4.892308	2.211856	1.087179	0.557583
Random Forest	1.533534	1.238360	0.801462	0.861321
Gradient Boosting	1.478839	1.216075	0.757480	0.866267
XGBoost	1.849176	1.359844	0.813819	0.832777

The below picture depicts the relationship between the actual student grade and the predicted values by the regularised lasso regression model in form of scatter plot. It is identified as the best model in this study. Each green points indicates the student actual and predicted values, the red line is called as best line of fit which should correctly matches the actual prediction and prediction made by the model.

Fig:1 lasso Regression actual and predicted values



4.2 Cross-Validation Stability

Table :2 cross validation with different folds

K-Folds	Model	RMSE Mean	RMSE Std	R2 Mean	R2 Std
2	Linear Regression	1.359220	0.055279	0.822203	0.002171
	Ridge Regression	1.358644	0.055462	0.822355	0.002223
	Lasso Regression	**1.335812**	0.056941	**0.828293**	0.002770
	Decision Tree	1.849987	0.125949	0.664679	0.067969
	Random Forest	1.285882	0.046877	0.840831	0.000600

	Gradient Boosting	1.332366	0.017152	0.828756	0.007431
	XGBoost	1.464365	0.003438	0.792742	0.015282
3	Linear Regression	1.302982	0.240080	0.835854	0.039603
	Ridge Regression	1.302047	0.240701	0.836091	0.039712
	Lasso Regression	**1.286389**	0.246498	**0.839865**	0.041084
	Decision Tree	2.028734	0.114013	0.599856	0.032982
	Random Forest	1.273114	0.181134	0.843537	0.025239
	Gradient Boosting	1.325885	0.230461	0.830308	0.036931
	XGBoost	1.421674	0.263546	0.804913	0.046336
4	Linear Regression	1.277717	0.245807	0.841496	0.031798
	Ridge Regression	1.277321	0.246593	0.841598	0.031974
	Lasso Regression	**1.268223**	0.248106	**0.843789**	0.032546
	Decision Tree	1.839589	0.204240	0.662137	0.065095
	Random Forest	1.252653	0.163084	0.846882	0.013561
	Gradient Boosting	1.302038	0.181201	0.835067	0.014347
	XGBoost	1.316464	0.209128	0.831037	0.026489
5	Linear Regression	1.284212	0.280960	0.839419	0.034048
	Ridge Regression	1.283782	0.281618	0.839535	0.034195
	Lasso Regression	**1.274306**	0.284174	**0.841850**	0.035001
	Decision Tree	1.880055	0.309294	0.654061	0.051424
	Random Forest	1.277723	0.254060	0.841073	0.027084
	Gradient Boosting	1.297490	0.306248	0.835066	0.042756
	XGBoost	1.382659	0.316975	0.813883	0.043443
6	Linear Regression	**1.266648**	0.275606	**0.840932**	0.049435
	Ridge Regression	**1.266158**	0.275717	**0.841066**	0.049407
	Lasso Regression	**1.257119**	0.278292	**0.843318**	0.049806
	Decision Tree	1.696306	0.325286	0.718601	0.066361
	Random Forest	1.230634	0.227812	0.849583	0.041397

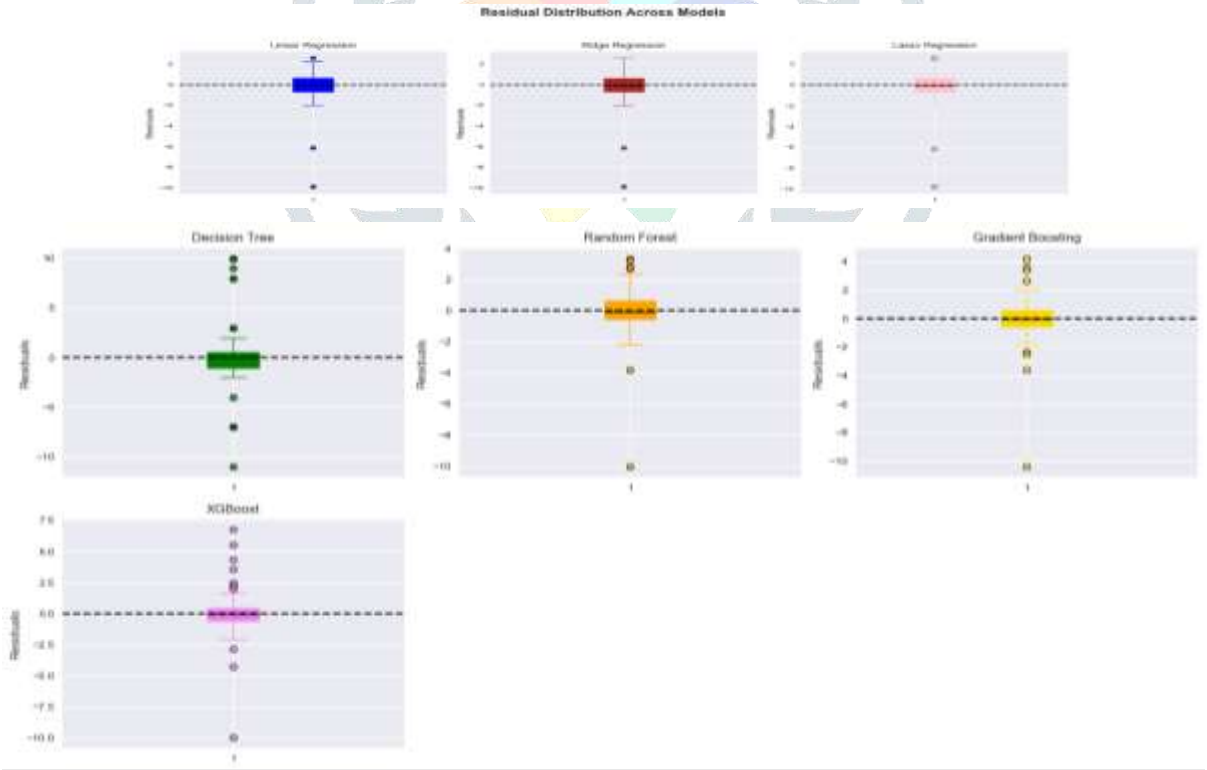
	Gradient Boosting	1.275466	0.253194	0.837877	0.047743
	XGBoost	1.318270	0.237786	0.828395	0.045144

The table 2 shows the cross validation stability for different k values has been tested for different regression models to find the consistent pattern in model performance. **Random Forest and Lasso Regression constantly attained the lowest RMSE values and the highest R^2 scores across different splits. The linear ,lasso and ridge regression produced constant values when split is k=6. When the split is k=2 and k=3 the lasso regression yields a lower R value when compared with splits 4, 5 and 6.** In contrast, the **Decision Tree showed significantly higher error and unstable R^2 values**, indicating poor reliability regardless of fold choice. XGBoost achieved moderate performance but with slightly higher variability than Random Forest and Gradient Boosting Finally, **regularized linear models (Lasso)** are the most stable and effective predictors of student performance across varying validation strategies.

4.3 Residual Analysis

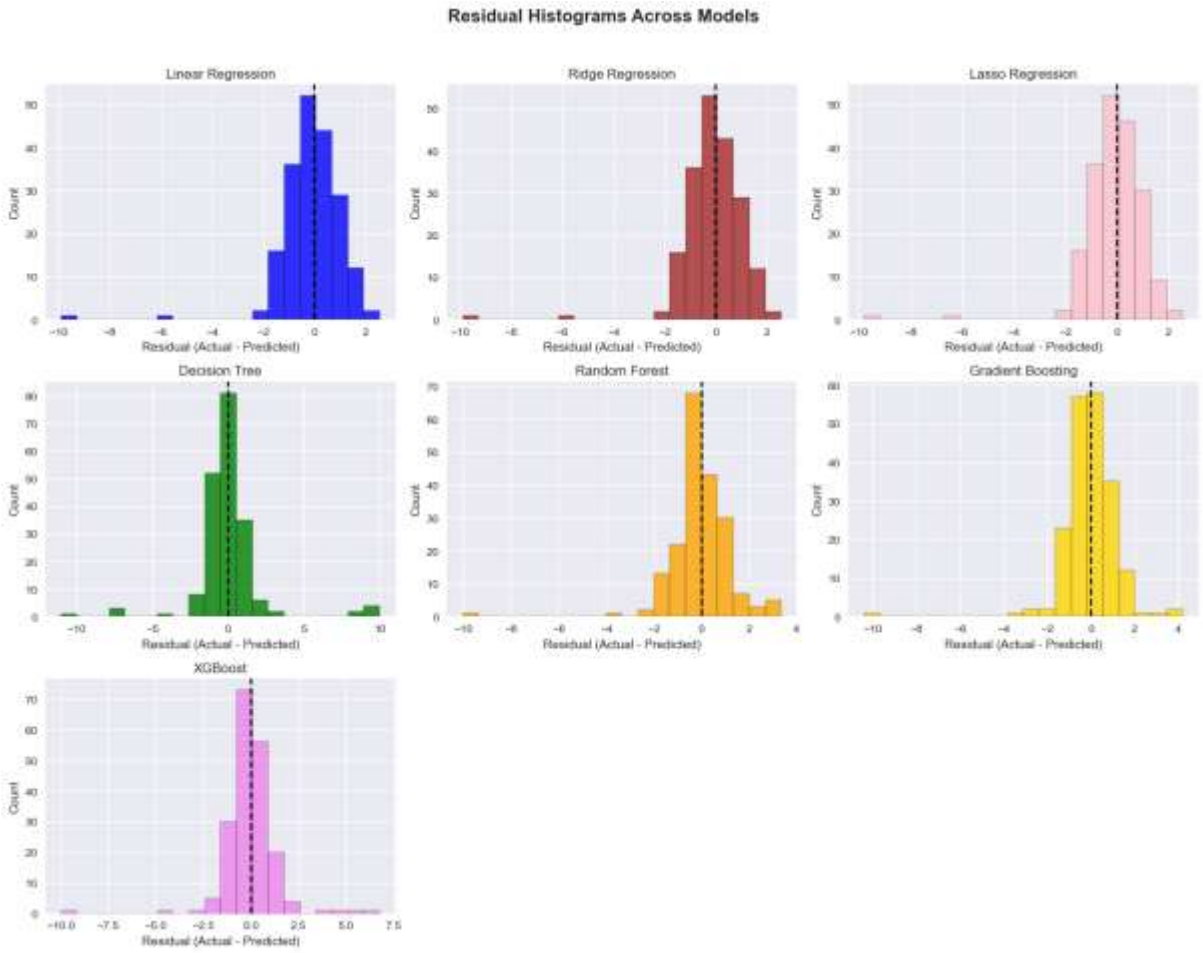
The residuals are the differences between actual and predicted values are examined through scatter plots, histograms, and boxplots. The visual representation shows the accuracy of predictions and highlights the systematic deviations and biases. The residual plot is demonstrated in below figure 3. It clearly shows that linear regression, ridge regression and lasso regression excels symmetric distribution around zero and indicates a good fit and minimal bias. The ensemble models such as Random Forest, Gradient Boosting and XG Boost shows wider spread still continues to maintain balanced plots of residual patterns indicates capable of capturing the complex relationship. On the other hand, the Decision Tree shows a very wide range of variations and extreme outliers shows its instability with the dataset. On the whole, regularized linear models outperforms in predicting the student marks.

Fig:3 Residual Distribution



4.4 Residual Histogram Analysis

Fig:4 Residual Histogram



The above figure 4 shows the histogram for different regression models. It helps to know the further insights for model performance. The linear regression and regularization models such as Linear, Lasso and Ridge show a narrow distribution and centered in zero. It shows that errors are very minimal and unbiased. Among these algorithms Lasso regression have a compact spread shows its effectiveness in error reduction through regularization. The residual histogram analysis clearly shows that Linear Regression produces relatively small and consistent errors, with most residuals narrowed within the range of ±2. This indicates that the model is stable, unbiased, and generalizes well to unseen data. In contrast, the Decision Tree model exhibits residuals extending up to ±5, reflecting larger deviations between actual and predicted values.

The Decision tree have wider spread and also have irregular distribution which shows extreme instability and overfitting. whereas **ensemble methods such as Random Forest and Gradient Boosting** maintain tighter, bell-shaped distributions, reflecting robust generalization. **XGBoost also yields a concentrated distribution** but with slightly heavier tails compared to Random Forest and Gradient Boosting, pointing to occasional larger errors. Overall, from the histograms it confirms that **regularized regression models and ensemble approaches achieve the most reliable predictions**, while **Decision Tree suffers from large residual variance**.

V. CONCLUSION

The lasso regression shows the best predictive accuracy (R^2 = 0.874, RMSE = 1.18, MAE = 0.77) highest R value and lowest MAE which out performs all other regression models. The model tends to produce same and consistent results even when tested with different k-fold values. The residual distribution graph shows regression models such as linear regression, lasso regression and ridge regression maintains residuals around zero which indicates a good fit and stability of models whereas decision tree has wider residual spread and contain extreme outliers. In contrast Decision tree and XGBoost showed poor performance with lowest R^2 value of 0.56, it proves the issue of model overfitting. Lasso regression yields a constant stability among all other linear models whereas Random Forest demonstrated highest reliability R^2 = 0.85 with the lowest RMSE variance while tested with cross validation

stability. Over all the result suggest that lasso regression out performs the linear models and Random Forest is most stable and dependent model. These insights are important of using the regularization in educational data analytics. It ensures the consistent and accuracy in predicting the student outcomes.

REFERENCES

[1] K. Yan, "Student Performance Prediction Using XGBoost Method from A Macro Perspective," *Proc. CDS 2021*, pp. 1–6, 2021, doi: 10.1109/CDS52072.2021.00084.

[2] P. Kumar, "Evaluating Machine Learning Algorithms for Enhanced Prediction of Student Academic Performance," *Journal Article*, 2025. doi: 10.21203/rs.3.rs-5730131/v1.

[3] G. Airlangga, "A Comparative Analysis of Machine Learning Models for Predicting Student Performance: Evaluating the Impact of Stacking and Traditional Methods," *Brilliance*, vol. 4, no. 2, Oct. 2024, doi: 10.47709/brilliance.v4i2.4669.

[4] E. Deniz, "Evaluating the Effectiveness of Machine Learning Models in Predicting Student Academic Achievement," *Journal Article*, Jul. 2024, doi: 10.69882/adba.cs.2024072.

[5] F. Zhou and N. Agarwal, "Student Performance Prediction Based on Decision Trees," *Journal of Research in Applied Mathematics*, vol. 10, no. 12, pp. 114–120, Dec. 2024, doi: 10.35629/0743-1012114120.

[6] R. Qureshi and P. S. Lokhande, "A Comprehensive Review of Machine Learning techniques used for Designing An Academic Result Predictor And Identifying The Multi-Dimensional Factors Affecting Student's Academic Results," *2024 IEEE Conference on Intelligent Data Analytics*, Nov. 2024, doi: 10.1109/idicaiei61867.2024.10842901.

[7] A. Abatal, A. Korchi, M. Mzili, T. Mzili, H. Khalouki and M. E. K. Billah, "A Comprehensive Evaluation of Machine Learning Techniques for Forecasting Student Academic Success," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 1, Nov. 2024, doi: 10.35882/jeeemi.v7i1.489.

[8] N. Santiketa, S. Chaikhan, U. Ninrutsirikun and N. Wattanakitrungroj, "Student Academic Performance Prediction using Machine Learning with Various Features and Scenarios," *IEEE International Conference on Software Engineering and Computing (ICSEC)*, Nov. 2024, doi: 10.1109/icsec62781.2024.10770729.

[9] B. Owaidat, "Exploring the accuracy and reliability of machine learning approaches for student performance," *Applied Computer Science*, vol. 20, no. 3, pp. 35–44, Sep. 2024, doi: 10.35784/acs-2024-29.

[10] G. Airlangga, "A Comparative Analysis of Machine Learning Models for Predicting Student Performance: Evaluating the Impact of Stacking and Traditional Methods," *Brilliance*, vol. 4, no. 2, Oct. 2024, doi: 10.47709/brilliance.v4i2.4669.