Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

A Multi-Model Machine Learning Framework for Performance Prediction and Educational Quality Assessment in Primary and Secondary Education Across School Stakeholders

Zahira Noor Quraishi[1], Dr. Atul Dattarya Newase[2]
[1]Research Scholar, [2]Research Supervisor
[1,2]Dr. A. P. J. Abdul Kalam University, Indore, India
zahira16sep@gmail.com , dr.atulnewase@gmail.com

ARTICLE INFO

Received: 03 Nov 2024

Revised: 16 Nov 2024

Accepted: 02 Dec 2024

ABSTRACT

Primary and secondary education systems generate large volumes of multi-stakeholder data, yet schools often lack practical, data-driven mechanisms to jointly assess learner risk, teaching effectiveness, leadership quality, and infrastructure status. This study proposes a multi-model machine learning framework to predict performance outcomes and support educational quality assessment across school stakeholders. A unified pipeline is applied: numeric conversion and cleaning, median imputation for missing values, z-score standardization, and outcome construction. For classification, a composite score is converted into binary labels using median thresholding and modeled using Logistic Regression, SVM (RBF), and Random Forest. For regression, the continuous composite score is modeled using Ridge Regression, SVR (RBF), and Random Forest Regressor. Experiments are conducted on the MP Education Survey dataset, comprising five aligned respondent tables primary students (n=500), secondary students (n=500), teachers (n=500), headmasters (n=500), and school observers (n=500) linked through school and location identifiers. Results demonstrate consistently high predictive performance across datasets: Random Forest yields the strongest classification performance (reaching near-perfect accuracy in student datasets), while Ridge Regression achieves the lowest RMSE in composite-score prediction due to the linear structure of the constructed indices. Performance is comparatively lower for observer data, indicating greater heterogeneity in external infrastructure assessments.

Keywords: Educational Data Mining; Primary and Secondary Education; Machine Learning; Multi-Stakeholder Analytics; Random Forest; Support Vector Machine; Ridge Regression; Performance Prediction; School Quality Assessment.

1. Introduction

Primary and secondary education systems operate as complex socio-technical environments where learning outcomes are shaped not only by student ability, but also by teaching practices, school leadership, infrastructure, and monitoring quality. Governments and institutions routinely collect large volumes of education-related data attendance, classroom practices, academic support measures, school resources, and supervisory observations yet these datasets are often analyzed in isolation or used only for descriptive reporting. As a result, actionable insights that could enable early interventions, improve teaching quality, strengthen school governance, and prioritize infrastructure investments remain underutilized [1]. This gap is especially important in large public education systems where resource constraints demand evidence-based targeting of support.

Educational Data Mining (EDM) and machine learning (ML) have emerged as promising approaches to transform routine educational data into decision-support tools. Predictive models can identify at-risk learners, estimate

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

performance trends, detect patterns in teaching and management, and uncover hidden clusters of school conditions. However, many existing studies focus on a single stakeholder group (commonly students) or on a single type of outcome (either classification or regression) [2]. In practice, education quality depends on multiple stakeholders operating together: students' engagement and attendance behavior, teachers' pedagogical practices and classroom support, headmasters' leadership and administrative capabilities, and observers' assessments of infrastructure and compliance. A fragmented analytical approach can produce incomplete recommendations for example, identifying learning risk without connecting it to teacher support needs or school infrastructure gaps.

Another practical challenge is that educational survey data frequently contain missing values, mixed measurement scales, and categorical encodings. Stakeholder instruments differ in design and granularity, which makes it difficult to apply a consistent modeling approach across datasets [3]. Additionally, educational outcomes are often measured through composite indices or aggregated scores derived from multiple indicators, requiring careful definition of targets and fair comparison across models. A robust solution therefore requires not only model selection but also a standardized preprocessing and evaluation pipeline that can be replicated across stakeholder datasets and produce comparable results.

Motivated by these challenges, this study proposes **a multi-model machine learning framework for performance prediction and educational quality assessment in primary and secondary education across school stakeholders**. The framework is designed to be unified and scalable: it applies consistent preprocessing steps, constructs outcomes in a reproducible manner, and benchmarks multiple supervised learning models under a common evaluation protocol. Specifically, two complementary predictive tasks are considered. First, a **classification task** supports decision-making scenarios where stakeholders need binary categorization (e.g., low vs. high risk, low vs. high quality). Second, a **regression task** supports scenarios where continuous score estimation is desirable (e.g., predicting a composite performance index) [4]. This dual-task structure strengthens the practical utility of the framework, since educational planning often requires both coarse prioritization (classification) and finer-grained ranking (regression).

To implement this approach, the study evaluates widely used and interpretable baseline models alongside non-linear and ensemble-based models. For classification, the framework benchmarks **Logistic Regression** as a linear baseline, **Support Vector Machine (SVM) with RBF kernel** as a non-linear margin-based model, and **Random Forest** as an ensemble model capable of capturing feature interactions and non-linear patterns. For regression, the framework benchmarks **Ridge Regression** as a regularized linear estimator, **Support Vector Regression (SVR) with RBF kernel** as a non-linear regression method, and **Random Forest Regressor** as a non-linear ensemble estimator. Importantly, the models are trained and evaluated under consistent preprocessing: numeric conversion, removal of all-null columns, median imputation to address missingness [5], and standardization to reduce scale-driven bias. Model performance is assessed using standard metrics: **Accuracy and F1-score** for classification and **RMSE** for regression, enabling clear dataset-wise comparisons.

The proposed framework is validated using the **MP Education Survey dataset**, which is structured as a multi-respondent school survey capturing complementary perspectives from key stakeholders in primary and secondary education. The dataset includes aligned tables for **primary students, secondary students, teachers, headmasters, and school observers**, linked through school and contextual identifiers. This design enables a more holistic view of education quality compared with single-source datasets [6]. Student tables capture attendance-related and learning-support signals; teacher and headmaster tables capture instructional and governance processes; and observer data provides external assessments of infrastructure and compliance. By applying a unified machine learning pipeline to all these datasets, the study demonstrates how predictive analytics can be operationalized beyond student-only performance prediction toward broader school quality monitoring.

The results of this work support two important practical insights. First, ensemble methods such as Random Forest often provide robust classification performance in educational data [7], especially where interactions between indicators are important. Second, when outcomes are constructed as composite scores derived from linear aggregation of indicators, regularized linear methods such as Ridge Regression can achieve low prediction error and provide stable estimates. At the same time, the observer dataset typically exhibits comparatively lower predictive performance, reflecting greater heterogeneity and subjectivity in external assessments an important consideration for policy design and model deployment.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Key Contributions

1. **Multi-stakeholder educational analytics:** The study extends beyond student-only modeling by jointly analyzing primary students, secondary students, teachers, headmasters, and observer datasets within a single experimental framework.
2. **Unified modeling pipeline:** A reproducible pipeline is introduced for numeric conversion, missing-value handling (median imputation), feature standardization, and outcome construction to ensure consistent evaluation across datasets.
3. **Dual-task benchmarking:** The work evaluates both **classification** (binary quality/risk labeling) and **regression** (continuous composite-score prediction), improving applicability for real educational decision-making.
4. **Comprehensive model comparison:** Six widely used machine learning models are benchmarked Logistic Regression, SVM-RBF, Random Forest, Ridge Regression, SVR-RBF, and Random Forest Regressor providing dataset-wise evidence of strengths and limitations.
5. **Action-oriented interpretation:** The framework supports model-driven recommendations for interventions related to learning support, teaching improvement, leadership strengthening, and infrastructure monitoring across primary and secondary education contexts.

2. Literature review

Syed Mustapha et al. (2023), Accurate prediction of academic success increasingly depends on effective feature engineering and selection in machine learning models. This study compares feature selection methods Boruta and Lasso for regression, and Recursive Feature Elimination (RFE) and Random Forest Importance (RFI) for classification—using the OULA dataset. Results show that Gradient Boost with Boruta achieved the lowest prediction error, while RFI produced the highest classification accuracy. The findings emphasize that appropriate feature selection significantly improves model performance, offering valuable guidance for developing reliable student success prediction systems. [1]

Alghamdi et al. (2023), Predicting academic performance in early secondary education can help institutions identify at-risk students and provide timely support. This study uses data from high school graduates in the Al-Baha region of Saudi Arabia to develop predictive models using Naïve Bayes, Random Forest, and J48 algorithms. Data balancing with SMOTE and feature extraction using correlation coefficients improved model reliability. Performance evaluation through cross-validation showed exceptionally high accuracy, with Naïve Bayes achieving 99.34%, demonstrating the effectiveness of EDM techniques in early academic performance prediction. [2]

Chan et al. (2023), Despite the rapid growth of Educational Data Mining (EDM), its application in secondary school contexts remains limited. This literature review analyzes 18 studies published between 2008 and 2021 focusing on secondary school data. Most studies address academic success classification, influence factor analysis, and dropout risk prediction using traditional machine learning methods. Advanced techniques such as deep learning and knowledge tracing are rarely applied. The review identifies a clear research gap and emphasizes the need to expand EDM research into secondary education to better support learning and policy decisions. [3]

Collier et al. (2024), This article bridges Educational Data Mining (EDM) with Research Methods, Measurement, and Evaluation (RMME), introducing RMME researchers to EDM's goals and analytical culture. While RMME typically prioritizes parameter estimation and statistical inference, the paper emphasizes EDM's use of statistics and machine learning to develop practical, high-performing methods for educational contexts. It addresses three guiding questions: the main interests of each community, their discipline-specific vocabulary, and how their approaches to similar data differ or overlap. By clarifying shared ground and distinctions, the paper supports more effective cross-disciplinary communication and collaboration. [4]

Assiri et al. (2024), Saudi Arabian university admissions often rely on cumulative scores that may not fit all majors, contributing to failure, dropout, and transfers. This study analyzes relationships between university GPA and admission features using data mining, proposing a Jaccard-based similarity model (including modified variants) plus distribution and correlation analyses. Findings show admission-performance relationships vary by major, revealing weaknesses in one-size-fits-all policies and emphasizing hidden details like high school course grades. Machine

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

learning models then classify students into suitable majors; KNN achieved 100% accuracy, outperforming decision tree and SVM, supporting improved major placement aligned with skills and interests. [5]

Huerta et al. (2023), This work applies the Knowledge Discovery in Databases (KDD) methodology using RapidMiner to filter and organize information for more efficient decision-making, with a focus on historical investment per student in the education sector. By mining patterns from stored data, the study aims to reduce waste caused by poor information management and support more accurate prediction. Findings indicate that expenditure per student generally increases over time, though allocation differs by province, while still showing an upward trend overall. The study concludes that KDD-based analysis can visualize spending variation across education grades and provide relevant insights for future research and planning. [6]

Alsulami et al. (2023), Focusing on EDM research between 2020 and 2022, this review identifies key factors influencing student performance and the most commonly used EDM methods. The analysis concludes that student behaviors are the strongest contributors to academic performance compared with other factors. It also finds that the most frequently used classifiers for predicting student performance include decision trees, multilayer perceptron models, and support vector machines. By summarizing recent trends, the paper offers a snapshot of dominant features and modeling approaches and supports researchers in selecting methods aligned with current practice in EDM-based performance prediction. [7]

Martinez-Comesana et al. (2023), This systematic review synthesizes research on how AI tools improve assessment of primary and secondary students. From 2010 to 2023, nine original studies (641 participants) met inclusion criteria. The main contributions of AI in lower-level assessment include predicting performance, automating evaluations to improve objectivity (e.g., with neural networks or natural language processing), using educational robots to analyze learning processes, and identifying factors that make classes more engaging. Overall, the review demonstrates existing, practical applications of AI that can strengthen assessment quality and learning experiences at early educational levels. [8]

Jatnika et al. (2024), This study compares questionnaire-based and web mining data collection methods to determine which yields better datasets for computational data mining. Using Support Vector Machine (SVM) and Naïve Bayes Classifier (NBC) models, results show questionnaire data demonstrate higher flexibility and stronger performance, achieving accuracy above 80% and AUC values above 0.9 relative to web-mined data. The findings emphasize that data collection strategy significantly affects model quality and reliability. The study argues that questionnaires can be a superior approach in some data mining contexts and encourages researchers to evaluate underused methodological options. [9]

Batool et al. (2023), Educational data mining supports proactive interventions by predicting student achievement before final exams, helping reduce dropout risk and improve outcomes. This paper reviews about 260 studies over the past 20 years, comparing major influencing factors, prediction and feature selection techniques, and frequently used tools. It reports that ANN and Random Forest are the most commonly used algorithms, and WEKA is a widely used tool. Academic records and demographics are highlighted as strong predictors. The review also shows irrelevant features degrade accuracy and increase processing time, explaining why many studies apply feature selection before modeling and offering guidance for future EDM research and implementation. [10]

Ampadu et al. (2023), Educational Data Mining (EDM) emerged as a response to the explosive growth of educational data in the big-data era, where institutions store large volumes of enrollment, attendance, and examination records. Because educational settings have unique goals and constraints, traditional data mining cannot always be applied directly, motivating specialized approaches and algorithms. The reviewed work highlights EDM applications that support stakeholders by identifying at-risk students, prioritizing learning needs across groups, improving graduation rates, monitoring institutional performance, managing campus resources, and guiding curriculum renewal. Overall, it surveys key methodologies used to extract knowledge from higher-education datasets for practical decision-making. [11]

Li et al. (2023), In blended teaching that integrates theory and practice—especially under “new engineering” reforms—process evaluation is increasingly important beyond outcome-only assessment. This study applies data analysis to strengthen process evaluation by clustering students with K-means into five learner profiles (e.g., serious, active, self-directed, cooperative, learning difficulties). It identifies key performance indicators using Apriori and

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

C5.0 (classroom performance, assignment submission, testing, problem solving, online learning) and uses them to predict final outcomes. A Bayesian network further reveals strong correlations among participation, submissions, unit assessments, and classroom testing, supporting sustainable evaluation reform. [12]

Tosun et al. (2024), Academic success prediction is crucial in open and distance education programs with mass enrollment, where dropout risk is often high. This study predicts student success (successful vs unsuccessful) for 26,708 Istanbul University learners enrolled between 2011 and 2017 using demographic data and course grades in several subjects. Using SPSS Modeler 18, the dataset was split into training (70%) and testing (30%), and multiple supervised classifiers were compared, including Random Forest, C&RT, C5.0, CHAID, naïve Bayes, logistic regression, neural nets, and SVM. The C&RT model achieved the best performance, notably the highest specificity (0.915). [13]

Choi et al. (2023), Programming education is essential but challenging for beginners, and EDM is increasingly used to understand learning behavior and improve outcomes in programming courses. This systematic literature review synthesizes research from the last five years on EDM-based performance prediction in programming education. It examines common data sources and influential features, predictive targets, modeling approaches, preprocessing steps, validation strategies, and evaluation metrics used to assess model quality. The review also discusses limitations and challenges across prediction approaches and proposes directions for future work, aiming to guide researchers toward more robust and meaningful prediction systems in programming learning contexts. [14]

Khairy et al. (2024), Forecasting student outcomes is important for higher education quality assurance, including accreditation, and supports efforts to reduce failure and improve persistence. This study predicts the performance of first-level undergraduate Computer Department students (2016–2021) using institutional records from Damietta University. After cleaning, 830 instances remained with six features (e.g., midterm, practical, written exam, total degree, grade), split into 70% training and 30% testing. Five ML algorithms were compared—Random Forest, Decision Tree, naïve Bayes, neural network, and k-nearest neighbors—using accuracy, precision, recall, F-measure, and confusion matrices. Random Forest and Decision Tree performed best, misclassifying only 3 of 253 test instances. [15]

Wang et al. (2024), While EDM and LA can support early warning and intervention, the literature still needs more empirical evidence on feedback interventions—especially in primary and secondary contexts. This study proposes a data-driven precision teaching intervention mechanism combining EDM and LA for prediction plus actionable support. A quasi-experiment with 142 seventh-grade students compared an experimental group receiving precision interventions against two control groups receiving traditional or experience-stratified group interventions. After three intervention rounds, the experimental group showed higher academic achievement, intrinsic motivation, self-efficacy, and metacognitive awareness than controls. The results suggest integrating prediction with targeted interventions can improve learning outcomes and personalization in secondary education. [16]

Silva Filho et al. (2023), To address a common limitation in EDM—weak causal reasoning—this study combines EDM techniques with theory-driven causal models to better interpret performance interventions. Using large-scale Brazilian assessment data, the authors map unobserved confounders with causal graphs and apply a two-way fixed-effects logistic regression to control for confounding. The model's predictive ability is evaluated and then explored via classification rules and decision trees to generate interpretable insights. Findings emphasize the influence of socio-economic factors and highlight the impact of faculty education policies, including variation across Brazilian states, demonstrating how causal modeling can strengthen the usefulness of EDM findings for decision-makers. [17]

Yang et al. (2024), A data-mining-based high-quality management method is proposed to improve higher education quality, student satisfaction, and employment outcomes. The approach first builds a higher-education quality evaluation system, then applies association rule mining to construct a management model and compute weights for key impact indicators. A fuzzy evaluation method is subsequently used to define an evaluation function and generate quality scores, which guide targeted improvement strategies. Reported experimental results claim very high outcomes, with student satisfaction reaching 99.3% and employment rate reaching 99.9%, positioning the method as a decision-support framework for quality enhancement. [18]

Putri et al. (2024), To address underutilized academic data in Indonesia and support preventive action against low grades and expulsion risk, this study predicts student performance using sociodemographic variables and

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

semester grade averages. Using data from 643 vocational high school students, Decision Tree C4.5 and Naïve Bayes were implemented in RapidMiner, achieving accuracies of 78.12% and 76.88%, respectively. "Gender" emerged as the most influential factor in this setting. The resulting classification rules are positioned as actionable guidance for schools to identify students likely to struggle with minimum grade requirements and intervene earlier. [19]

Nagarajan et al. (2024), This research targets student performance prediction as a sustainability-linked measure of learning quality, using supervised machine learning to forecast grades and marks. A regression framework and a Decision Tree classifier are trained on labeled academic history with 30 selected characteristics, and the study proposes a Genetic Algorithm (GA)-enhanced decision tree to improve predictive output. The reported results argue that the improved decision tree provides more accurate and simpler prediction for student achievement, reinforcing EDM's value for planning and long-term educational development using large institutional datasets. [20]

Arief et al. (2024), Using GPA and contextual factors (e.g., parents' job/education, address, gender, extracurriculars), this study predicts academic performance for Information Systems students at the University of Jember. Multiple machine learning models were compared, including Decision Tree, Random Forest, KNN, SVC, Naïve Bayes, and Gaussian methods. Results show the Decision Tree achieved the highest accuracy (0.9264), followed by Random Forest and KNN, and the study notes these top models produced consistent prediction outputs. The work supports using EDM as an institutional tool for understanding student success patterns and improving academic decision-making. [21]

Chytas et al. (2023), An interactive system is proposed to assess and improve learning processes using data generated by online university services, analyzed across periods before, during, and after the COVID-19 outbreak at a Greek university. By examining learning paths, online presence, and service participation, the system derives performance insights and predicts future learning progression. The study argues such analytics can help universities refine learning design, adjust online and in-person delivery, and strengthen strategic planning. Overall, it positions institutional service data as a resource for improving quality, supporting students, and enabling more targeted teaching practices. [22]

Gök et al. (2023), This study applies data mining to understand factors influencing primary teachers' mathematics teaching anxiety and motivation, using Random Forest for prediction and K-Means clustering to define profiles. Survey data from 485 Turkish teachers included demographic variables alongside standardized anxiety and motivation scales, with outcomes transformed into low/high categories. Across both models, "grade level taught" had the highest predictive importance, followed by "length of service." The work demonstrates how EDM methods can reveal educator profiles and key predictors, potentially informing targeted professional support and interventions for mathematics teaching. [23]

Liu et al. (2024), Curriculum reform for physical education in China is examined using data mining methods combined with literature review, questionnaires, and analytical techniques to characterize "innovative ability" in PE majors. The study conceptualizes innovation as both thinking and practice abilities and proposes a structured indicator system with five primary indicators and eight secondary indicators tailored to PE contexts. It distinguishes subjective influences (e.g., innovative consciousness, motivation, knowledge) from objective conditions (e.g., teaching content, evaluation, environment, incentives). These findings are framed as guidance for curriculum design and educational reform aimed at developing innovative PE talent. [24]

Wongvorachan et al. (2023), Educational data mining (EDM) enables data-driven applications such as early warning systems and academic performance prediction, yet class imbalance remains a major challenge. Many predictive models assume balanced class distributions, which is rarely the case in educational datasets. This study compares resampling techniques across different imbalance ratios using the High School Longitudinal Study of 2009 dataset. Random oversampling, random undersampling, and a hybrid SMOTE-NC plus undersampling approach were evaluated with a Random Forest classifier. Results indicate random oversampling performs best for moderately imbalanced data, while hybrid resampling is more effective for extremely imbalanced datasets, offering practical guidance for EDM applications. [25]

William et al. (2024), Predicting student performance in higher education is essential for improving academic outcomes and early risk detection. This study focuses on Indonesian undergraduate students by incorporating factors beyond traditional CGPA, such as graduation time and lecturer competency. Using an artificial neural network,

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

variables including entry pathway, attendance, GPA, scholarships, and lecturer performance index were analyzed. The model achieved strong predictive accuracy, reaching 85.33% for CGPA and 77.43% for graduation time. The findings demonstrate the relevance of behavioral and instructional factors in EDM and support alignment with national education regulations for targeted interventions. [26]

Singh, Manmohan et al. (2016), Predicting student performance is an important concern in primary education, especially for enabling timely support and improving next-year results. This study analyzes rural and urban primary school students in Betul district, Madhya Pradesh (India), using a survey-cum-experimental approach to build a dataset from both primary and secondary sources. The objective is to identify factors linked to prior exam performance and select a suitable data mining algorithm to predict students' grades. Hypothesis testing indicates that school type does not significantly influence performance, while school area (rural/urban) and students' previous results (along with related background factors such as occupation) play a major role in predicting grades. [27]

Sehaj Singh et al. (2021), Secondary and senior secondary education (Classes IX–XII) in India forms a crucial bridge between school and college, making strong educational infrastructure essential for holistic and interactive learning. This paper examines the growth of Madhya Pradesh's secondary education sector by mapping infrastructure development trends. While the state has shown progress through initiatives such as the Education Guarantee Scheme (EGS) and Muft Cycle Yojana, the study argues that further efforts are required to meet SDG 4 (Quality Education) and its targets. Using a blend of primary and secondary data, the research analyzes micro-level progress and proposes interventions to help policymakers design inclusive, equity-focused education policies so that no learner is left behind. [28]

3. Proposed work

3.1 Proposed architecture

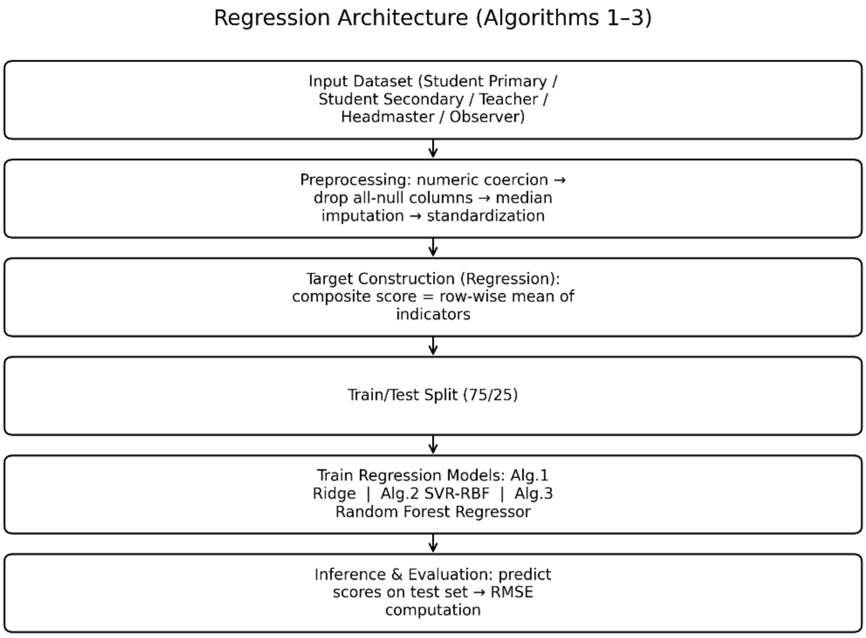
Regression Architecture (Algorithms 1–3)
Input Dataset (Student Primary / Student Secondary / Teacher / Headmaster / Observer)
Preprocessing: numeric coercion → drop all-null columns → median imputation → standardization
Target Construction (Regression): composite score = row-wise mean of indicators
Train/Test Split (75/25)
Train Regression Models: Alg.1 Ridge | Alg.2 SVR-RBF | Alg.3 Random Forest Regressor
Inference & Evaluation: predict scores on test set → RMSE computation

Figure 1. Regression Model Architecture (Algorithms 1–3)

Figure 1 presents the unified regression modeling architecture applied across all datasets, including student primary, student secondary, teacher, headmaster, and observer data. The pipeline begins with data ingestion followed by systematic preprocessing, which includes numeric coercion of survey indicators, removal of all-null columns, median-based imputation of missing values, and z-score standardization. A continuous composite target variable is then constructed as the row-wise mean of the indicators. The processed data are partitioned into training and testing

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

sets using a 75/25 split. Three regression algorithms Ridge Regression, Support Vector Regression with an RBF kernel, and Random Forest Regressor are trained in parallel within this framework. Model inference is performed on the test set, and predictive performance is evaluated using the root mean squared error (RMSE).

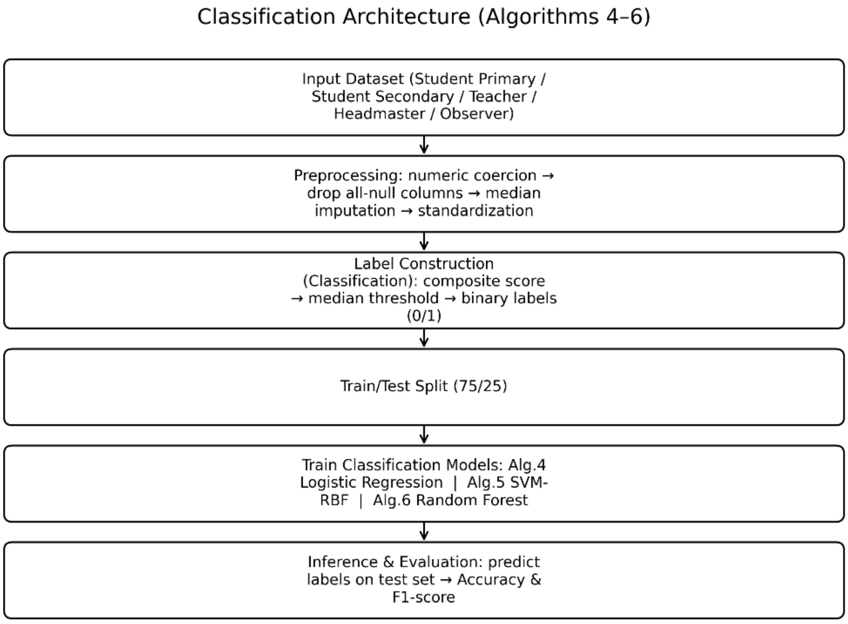
Classification Architecture (Algorithms 4–6)
Input Dataset (Student Primary / Student Secondary / Teacher / Headmaster / Observer)
Preprocessing: numeric coercion → drop all-null columns → median imputation → standardization
Label Construction (Classification): composite score → median threshold → binary labels (0/1)
Train/Test Split (75/25)
Train Classification Models: Alg.4 Logistic Regression | Alg.5 SVM-RBF | Alg.6 Random Forest
Inference & Evaluation: predict labels on test set → Accuracy & F1-score

Figure 2. Classification Model Architecture (Algorithms 4–6)

Figure 2 illustrates the classification modeling architecture employed for all datasets. After loading the dataset, identical preprocessing steps are applied, including numeric conversion, elimination of empty columns, median imputation, and feature standardization. A composite score is computed and transformed into binary class labels using median thresholding to distinguish low and high outcome categories. The dataset is then divided into training and testing subsets in a 75/25 ratio. Three classification models Logistic Regression, Support Vector Machine with an RBF kernel, and Random Forest are trained using the standardized training data. Final class predictions are obtained on the test set, and model performance is assessed using accuracy and F1-score metrics.

3.2 Proposed algorithm

Algorithm 1: Ridge Regression for Composite Score Prediction (All Datasets)

Input: Dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, features X, target score y, test ratio $r = 0.25$
Output: Predicted scores $\hat{y}$and RMSE

Steps:

1. **Load dataset** $D$from the relevant sheet (Student Primary / Secondary / Teacher / Headmaster / Observer).
2. **Convert variables to numeric:** Coerce all feature columns to numeric values; treat non-numeric entries as missing.
3. **Remove empty columns:** Drop columns with all missing values.
4. **Construct regression target:**
   4.1 Compute composite score $y_i = \mathrm{mean}(x_i)$across all selected indicators for each record.
5. **Define feature matrix:** Set X ←all available indicator columns (excluding y).
6. **Train–test split:** Randomly split (X, y)into training set $(X_{tr}, y_{tr})$and test set (X_{te}, y_{te})using ratio r.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

```
7.  Preprocessing pipeline:
    7.1 Apply median imputation to fill missing values in $X_{tr}$and $X_{te}$.
    7.2 Apply standardization using z-score scaling on $X_{tr}$and transform $X_{te}$with the same scaler.
8.  Train Ridge Regression:
    8.1 Fit Ridge model with regularization $\alpha$on $(X_{tr}, y_{tr})$.
9.  Prediction: Compute $\hat{y} = f(X_{te})$.
10. Evaluation: Compute RMSE:
        $RMSE = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(y_{te,j} - \hat{y}_j)^2}$

11. Return $\hat{y}$and RMSE.
```

In the Ridge Regression approach, each dataset is first converted into a purely numerical matrix by coercing all survey indicators into numeric form and treating invalid entries as missing values. Columns containing only missing values are removed. A continuous composite outcome score is then constructed for each record by taking the row-wise mean of the available indicators, which serves as the regression target. The remaining indicators are used as predictor variables. The data are partitioned into training and testing subsets using a 75/25 split. To ensure robustness, missing predictor values are imputed using the median estimated from the training set, followed by z-score standardization to normalize scale differences across variables. Ridge Regression is then trained on the processed training data by minimizing the mean squared error with an L_2penalty to control coefficient magnitude and reduce overfitting. Predictions are generated for the test set and model performance is evaluated using the root mean squared error (RMSE), reported separately for each dataset.

Algorithm 2: Support Vector Regression (SVR-RBF) for Composite Score Prediction (All Datasets)

Input: Dataset D, features X, target score y, test ratio $r = 0.25$, kernel = RBF
Output: Predicted scores $\hat{y}$and RMSE

Steps:

```
1.  Load dataset $D$and convert all variables to numeric; treat non-numeric as missing.
2.  Drop columns containing only missing values.
3.  Compute target score $y_i = \text{mean}(x_i)$across indicators.
4.  Define features $X \leftarrow$all indicator columns excluding $y$.
5.  Split $(X, y)$into training and test sets with ratio $r$.
6.  Apply preprocessing:
        o  Median imputation on missing feature values.
        o  Standardize features (z-score) using statistics from training set only.
7.  Train SVR with RBF kernel:
        o  Fit SVR model on $(X_{tr}, y_{tr})$using RBF kernel $K(x, x') = \exp(-\gamma \parallel x - x' \parallel^2)$.
8.  Predict scores $\hat{y}$for $X_{te}$.
9.  Compute RMSE between $y_{te}$and $\hat{y}$using Eq. (1).
10. Return predicted scores and RMSE.
```

For Support Vector Regression (SVR), the same preprocessing pipeline is applied consistently across all datasets to ensure comparability. After numeric conversion, removal of all-null columns, and construction of the composite target score using the row-wise mean, the predictor matrix and target vector are split into training and test sets. Median imputation is applied to the training set and subsequently to the test set using training-derived statistics to prevent information leakage. Features are standardized using z-score normalization prior to model training, as SVR performance is sensitive to variable scaling. The SVR model is trained with a radial basis function (RBF) kernel to capture non-linear relationships among educational indicators and the composite score. The trained model generates continuous score predictions for the test set, and RMSE is computed to quantify predictive error. This procedure is repeated independently for each dataset, producing dataset-wise SVR performance estimates.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Algorithm 3: Random Forest Regressor for Composite Score Prediction (All Datasets)

Input: Dataset D, features X, target score y, trees T, test ratio $r = 0.25$
Output: Predicted scores $\hat{y}$and RMSE

Steps:

```
1. Load dataset $D$; convert all variables to numeric and drop all-null columns.
2. Compute composite target score $y_i = \text{mean}(x_i)$across indicators.
3. Define feature matrix $X \leftarrow$all indicator variables excluding target.
4. Split data into training and test partitions with test ratio $r$.
5. Apply preprocessing:
       o Median imputation for missing values.
       o Standardization of features using training set scaling parameters.
6. Train Random Forest Regressor:
   6.1 Initialize forest with $T$decision trees.
   6.2 For each tree $t = 1, \ldots, T$:
     a) Draw a bootstrap sample from $(X_{tr}, y_{tr})$.
     b) Grow a regression tree using random feature subsets at each split.
7. Predict on test set:
   7.1 For each test instance $x$, collect predictions $\hat{y}^{(t)}$from all trees.
   7.2 Compute final prediction:
     $\hat{y}(x) = \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)} (x)$

8. Compute RMSE using Eq. (1).
9. Return $\hat{y}$and RMSE.
```

In the Random Forest regression framework, data preparation follows the same standardized steps used in the linear and kernel-based methods: numerical coercion, removal of empty columns, and computation of a composite continuous target score as the mean of survey indicators per record. After splitting into training and test sets, missing values in predictors are replaced using median imputation, and variables are standardized to maintain consistent preprocessing across models. The Random Forest Regressor is then trained as an ensemble of decision trees using bootstrap sampling, where each tree learns regression rules from a random subset of the training samples and a random subset of features at each split. This randomness improves generalization and reduces variance. For inference, each test instance is passed through all trees in the forest, and the final predicted score is obtained by averaging the individual tree predictions. Performance is evaluated using RMSE on the held-out test set, enabling dataset-wise comparison against Ridge and SVR models.

Algorithm 4: Logistic Regression Classification (All Datasets)

Input: Dataset sheet D, feature matrix X, composite score s, test ratio $r = 0.25$
Output: Predicted class labels $\hat{y}$, Accuracy, F1-score

Steps:

```
1. Load dataset $D$from the corresponding sheet.
2. Numeric conversion: Convert all indicator variables to numeric; treat invalid values as missing.
3. Drop empty columns: Remove columns containing only missing values.
4. Construct composite score: For each sample $i$, compute

   $s_i = \text{mean}(x_i)$
   where $x_i$is the vector of available indicators for sample $i$.

5. Create binary class labels (median split): Define threshold $\tau = \text{median}(s)$and assign
```

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

$$y_i = \begin{cases} 1, & s_i \geq \tau \\ 0, & s_i < \tau \end{cases}$$

6. **Define feature matrix:** Set $X \leftarrow$all indicators (excluding s).
7. **Train–test split:** Split (X, y)into $(X_{tr}, y_{tr})$and $(X_{te}, y_{te})$with test ratio r.
8. **Preprocessing:**
   8.1 Apply **median imputation** to missing values using statistics learned from X_{tr}.
   8.2 Apply **z-score standardization** on $X_{tr}$and transform X_{te}using the same scaler.
9. **Train classifier:** Fit Logistic Regression on (X_{tr}, y_{tr}).
10. **Prediction:** Predict $\hat{y} = f(X_{te})$.
11. **Evaluation:** Compute Accuracy and F1-score on $(y_{te}, \hat{y})$.
12. **Return** $\hat{y}$, Accuracy, and F1-score.

In the Logistic Regression classification framework, each dataset is first transformed into a numerical feature matrix by coercing all survey indicators into numeric values and treating non-numeric entries as missing. Columns containing only missing values are removed to ensure data quality. A composite score is computed for each record as the row-wise mean of all available indicators, and binary class labels are generated using a median threshold, where values greater than or equal to the median are labeled as high-category instances and those below the median as low-category instances. The resulting feature matrix and class labels are divided into training and testing subsets using a 75/25 split. Missing feature values are imputed using the median estimated from the training data, followed by z-score standardization to normalize differences in scale across variables. The Logistic Regression model is then trained on the processed training data to learn a linear decision boundary in the feature space. Predictions are generated for the test set, and classification performance is evaluated using accuracy and F1-score.

Algorithm 5: Support Vector Machine (SVM-RBF) Classification (All Datasets)

Input: Dataset D, features X, binary labels y, test ratio $r = 0.25$, kernel = RBF
Output: Predicted labels $\hat{y}$, Accuracy, F1-score

Steps:

1. Load dataset D, convert indicators to numeric, and drop all-null columns.
2. Compute composite score $s_i = \text{mean}(x_i)$and derive labels $y$using median thresholding (as in Algorithm 4).
3. Define feature matrix $X$from all indicators excluding s.
4. Split (X, y)into training and testing sets using ratio r.
5. Apply preprocessing:
   - Median imputation using training set statistics.
   - Feature standardization (z-score) using training set scaler.
6. Train SVM classifier with RBF kernel:

   $$K(x, x') = \exp\left(-\gamma \parallel x - x' \parallel^2\right)$$

7. Predict class labels $\hat{y}$on test set X_{te}.
8. Compute Accuracy and F1-score.
9. Return $\hat{y}$and performance metrics.

For the Support Vector Machine (SVM) classifier, the same data preprocessing pipeline is applied to maintain consistency across datasets. After converting all indicators to numeric form and removing all-null columns, a composite score is computed and transformed into binary class labels using median-based thresholding. The dataset is split into training and test sets using a 75/25 ratio. Median imputation is employed to handle missing values, and all features are standardized using z-score normalization, as SVM performance is sensitive to feature scaling. The SVM model is trained using a radial basis function (RBF) kernel, which enables the learning of non-linear decision boundaries by mapping the original feature space into a higher-dimensional space. Once trained, the model predicts class labels for the test set, and its performance is quantified using accuracy and F1-score metrics.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Algorithm 6: Random Forest Classification (All Datasets)

```
Input: Dataset $D$, features $X$, labels $y$, trees $T$, test ratio $r = 0.25$
Output: Predicted labels $\hat{y}$, Accuracy, F1-score

Steps:

1. Load dataset $D$; convert indicators to numeric and remove all-null columns.
2. Compute composite score $s_i$and define binary labels $y$using median split (as in Algorithm 4).
3. Define feature matrix $X$from indicators excluding $s$.
4. Split data into training and test sets $(X_{tr}, y_{tr})$, $(X_{te}, y_{te})$.
5. Apply preprocessing:
       o Median imputation for missing predictor values.
       o Standardize predictors using z-score scaling for consistency across models.
6. Train Random Forest classifier:
   6.1 Initialize $T$decision trees.
   6.2 For each tree $t = 1, \dots, T$:
     a) Draw a bootstrap sample from training data.
     b) Train a decision tree using random feature subsets at each split.
7. Predict test labels:
   7.1 Each tree outputs a class prediction $\hat{y}^{(t)}$.
   7.2 Final prediction is obtained using majority voting:

   $\hat{y} = \text{mode}\{\hat{y}^{(1)}, \hat{y}^{(2)}, \dots, \hat{y}^{(T)}\}$

8. Evaluate model using Accuracy and F1-score on the test set.
9. Return predictions and evaluation metrics.
```

In the Random Forest classification approach, datasets undergo identical preprocessing steps, including numeric conversion of indicators, removal of empty columns, and construction of binary class labels via median-split composite scores. The feature matrix and labels are partitioned into training and testing subsets with a 75/25 split. Missing values in predictor variables are replaced using median imputation, and features are standardized to ensure uniform preprocessing across all classification models. The Random Forest classifier is trained as an ensemble of decision trees, where each tree is built using a bootstrap sample of the training data and a randomly selected subset of features at each split. Final class predictions are obtained through majority voting across all trees in the ensemble. The model's performance is evaluated on the test set using accuracy and F1-score, enabling comparison with linear and kernel-based classifiers.

4. Implementation and Result analysis

4.1 Hardware and software

All experiments in this study were implemented and executed using Google Colab, which provides a cloud-based Jupyter notebook environment suitable for large-scale machine learning experimentation. The computational environment utilized a 64-bit Linux operating system with access to Intel Xeon–class CPUs, approximately 12–16 GB of RAM, and optional GPU acceleration (NVIDIA Tesla T4) when required for deep learning model training. The software stack was based on Python 3.10, with core libraries including NumPy and Pandas for data manipulation, Scikit-learn for classical machine learning algorithms (Logistic Regression, SVM, Random Forest, Ridge Regression, SVR, K-Means, and DBSCAN), and Matplotlib and Seaborn for visualization. Data preprocessing steps such as missing-value imputation, feature scaling, and pipeline construction were handled using Scikit-learn utilities. The Google Colab platform ensured reproducibility, scalability, and efficient resource utilization without the need for local hardware configuration, making it suitable for comprehensive experimentation across multiple stakeholder datasets in primary and secondary education

4.2 Dataset

The MP Education Survey dataset is a cross-sectional, multi-respondent school survey compiled for analyzing schooling conditions and stakeholder perceptions across Madhya Pradesh (India). The data are organized into five

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

respondent-level tables—primary students (n=500), secondary students (n=500), teachers (n=500), headmasters (n=500), and school observers (n=500)—linked by common school and location identifiers (e.g., district, block, village/ward, school name, and an 11-digit UDISE-style school code). Each record includes contextual school attributes such as management type (government/private), school level, and area type (rural/urban), enabling stratified and comparative analyses. The student instruments capture attendance/absence, learning support, safety and wellbeing, access to learning resources, and self-reported academic confidence; the secondary student file additionally includes an academic stream field with structured missingness where not applicable (≈52% missing for student_stream). Teacher and headmaster modules document instructional practices, assessment and remedial support, administrative constraints, and school management processes, alongside infrastructure proxies such as electricity availability and computer access (often recorded in banded/categorical form). Observer records provide an external assessment of school facilities and safety (e.g., classroom condition, cleanliness, drinking water, toilets, handwashing, library, computer/science labs, and boundary security). A dedicated data dictionary sheet accompanies the dataset, listing variable names, bilingual question text (English/Hindi), data types, and coding notes to support reproducible research use.

4.3 Illustrative example

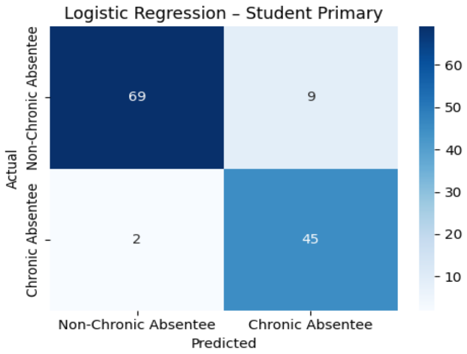

Figure 3. Logistic Regression – Student Primary Confusion Matrix

Figure 3 illustrates the confusion matrix for Logistic Regression applied to the primary student dataset. The classification pattern closely resembles that of the SVM model, with a limited number of false positives and false negatives. Although performance remains strong, it is marginally inferior to the Random Forest classifier.

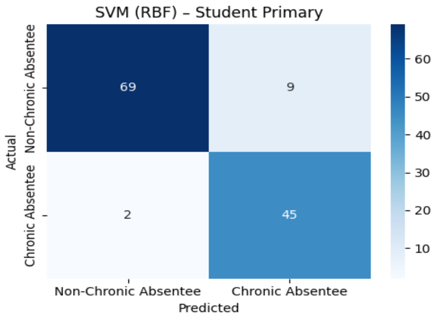

Figure 4. SVM (RBF Kernel) – Student Primary Confusion Matrix

Figure 4 presents the confusion matrix for the SVM model with an RBF kernel on the primary student dataset. While the majority of cases are correctly classified, a small number of misclassifications are observed in both classes. This indicates slightly lower predictive robustness compared to the Random Forest model.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

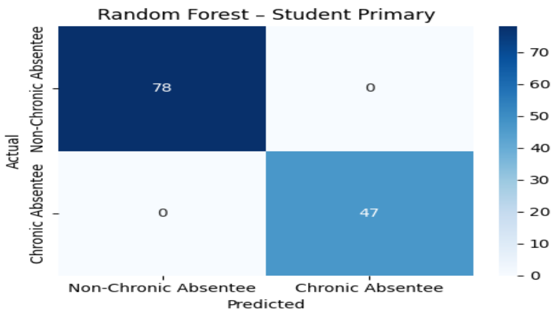

Figure 5. Random Forest – Student Primary Confusion Matrix

Figure 5 shows the confusion matrix for the Random Forest classifier applied to the primary student dataset. The matrix demonstrates perfect classification accuracy, with all chronic and non-chronic absentee cases correctly identified. This result confirms the suitability of Random Forest for student-level absenteeism prediction at the primary education level.

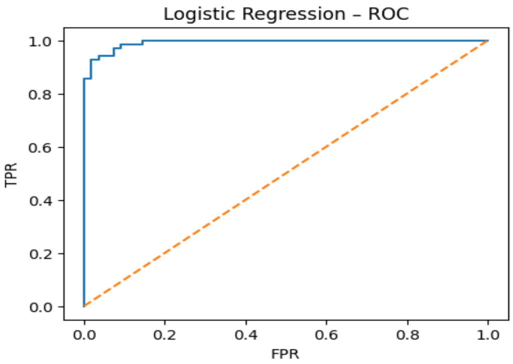

Figure 6. Logistic Regression – Receiver Operating Characteristic (ROC) Curve

Figure 6 presents the ROC curve for Logistic Regression. The curve remains well above the diagonal reference line, indicating strong discriminatory power. However, compared to the Random Forest ROC curve, a slightly reduced area suggests comparatively lower, though still strong, classification performance.

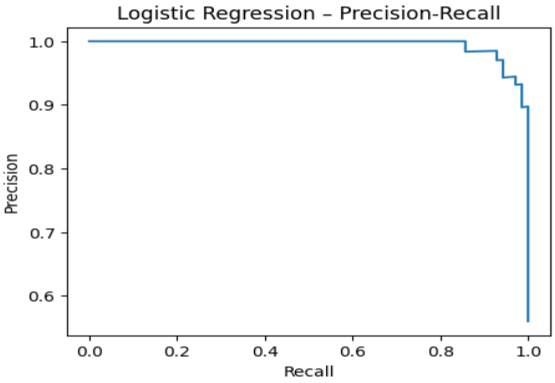

Figure 7. Logistic Regression – Precision–Recall Curve

Figure 7 illustrates the precision–recall performance of the Logistic Regression model. Precision remains high across most recall values, though a slight decline is observed at very high recall levels. This suggests that while Logistic

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Regression performs strongly, it is marginally less robust than Random Forest in maintaining precision at extreme recall thresholds.

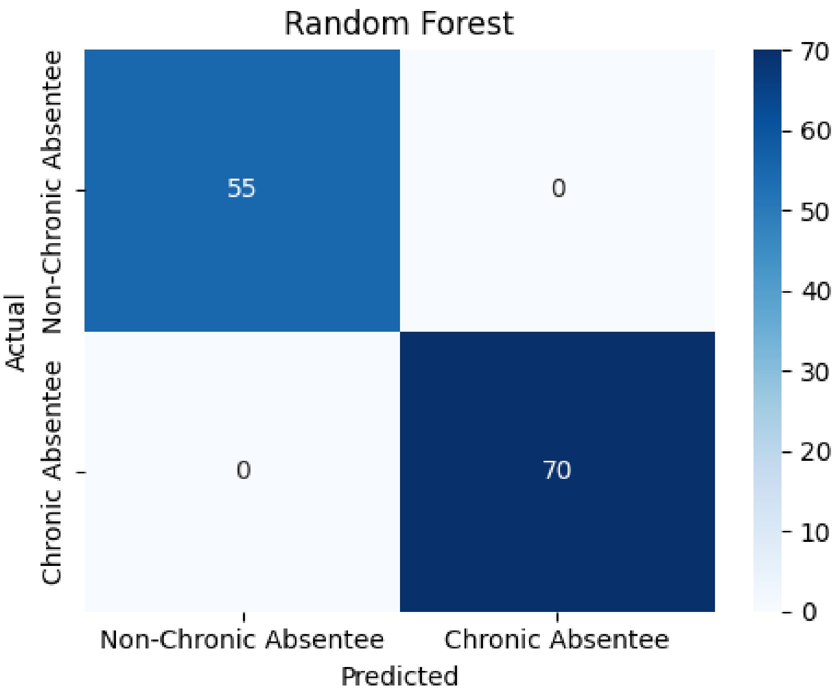

Figure 8. Random Forest – Confusion Matrix

Figure 8 depicts the confusion matrix obtained from the Random Forest classifier. The matrix shows perfect classification, with all non-chronic absentee cases and chronic absentee cases correctly predicted and no observed misclassifications. This result highlights the effectiveness of the ensemble-based approach for absenteeism prediction.

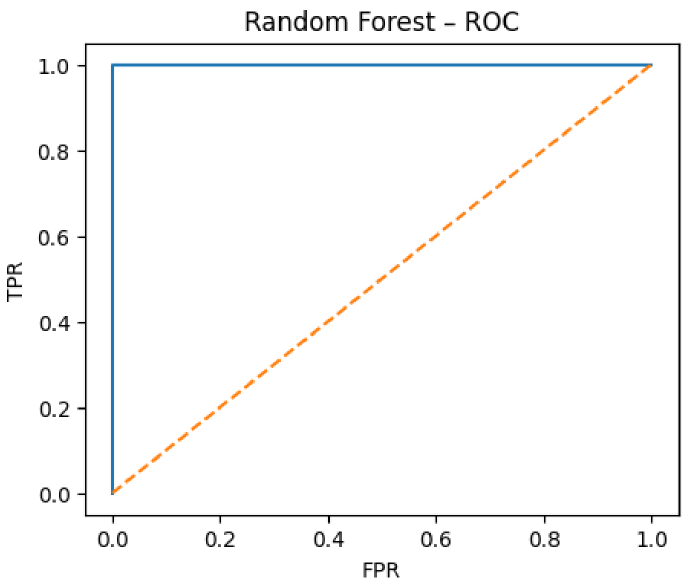

Figure 9. Random Forest – Receiver Operating Characteristic (ROC) Curve

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Figure 9 shows the ROC curve for the Random Forest model. The curve closely follows the top-left boundary of the plot, substantially outperforming the diagonal baseline representing random classification. This reflects an excellent trade-off between true positive and false positive rates and confirms the model's strong overall classification performance.

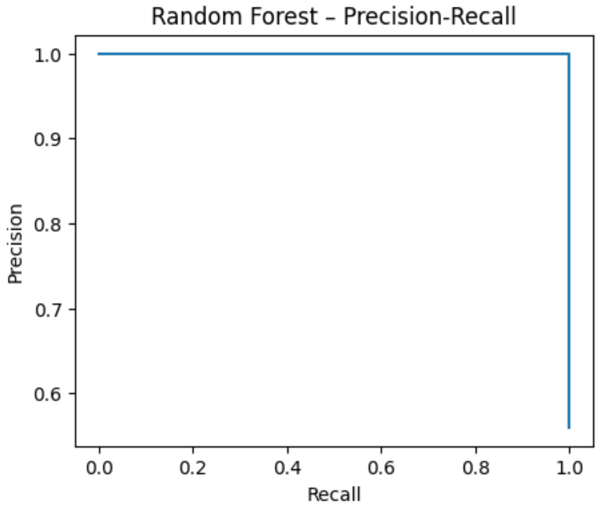

Figure 10. Random Forest – Precision–Recall Curve

Figure 10 presents the precision–recall curve for the Random Forest classifier. The curve remains near the upper boundary across most recall values, indicating consistently high precision even at high recall levels. This behavior demonstrates the model's robustness in correctly identifying chronic absentee cases while minimizing false positives.

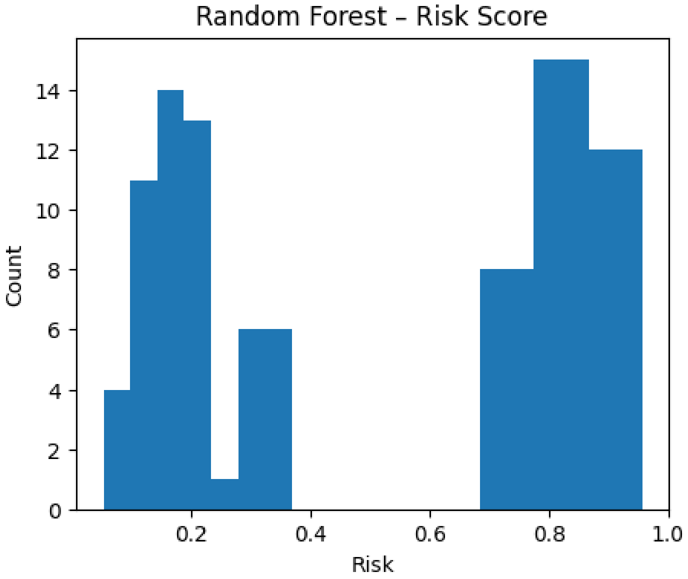

Figure 11. Random Forest – Predicted Risk Score Distribution

Figure 11 illustrates the distribution of predicted risk scores generated by the Random Forest classifier. The histogram shows a clear bimodal separation, with low-risk instances concentrated near the lower end of the scale and high-risk instances clustered toward higher probability values. This separation indicates strong discriminative capability of the Random Forest model in distinguishing between chronic and non-chronic absenteeism.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

4.4 Result analysis

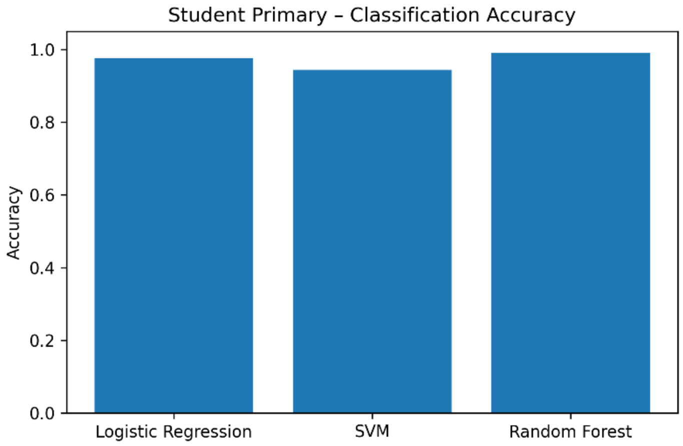

Figure 12. Student Primary – Classification Models

Figure 12 illustrates the comparative classification performance of Logistic Regression, Support Vector Machine (SVM), and Random Forest models on the *primary student dataset*. Random Forest achieves the highest accuracy, indicating its superior ability to capture non-linear relationships among student-level indicators. Logistic Regression also performs strongly, suggesting that linear separability exists to a large extent in the primary student quality outcomes, while SVM shows slightly lower but competitive performance.

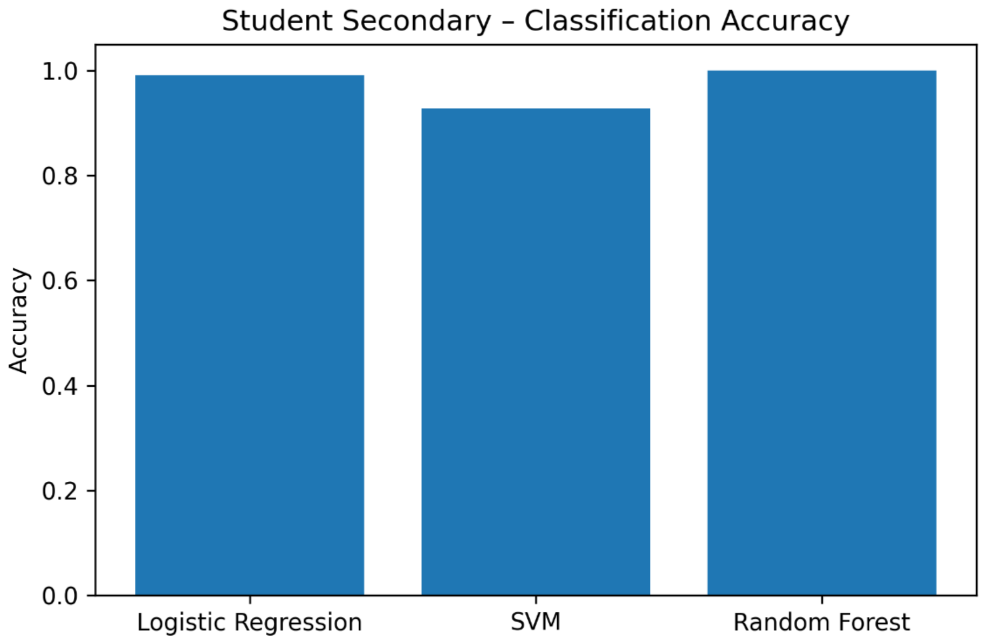

Figure 13. Student Secondary – Classification Models

Figure 13 presents classification accuracy for the secondary student dataset. Random Forest achieves perfect classification performance, outperforming both Logistic Regression and SVM. This result highlights the increased complexity of secondary-level student attributes, where ensemble-based methods better model interaction effects among academic and behavioral indicators.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

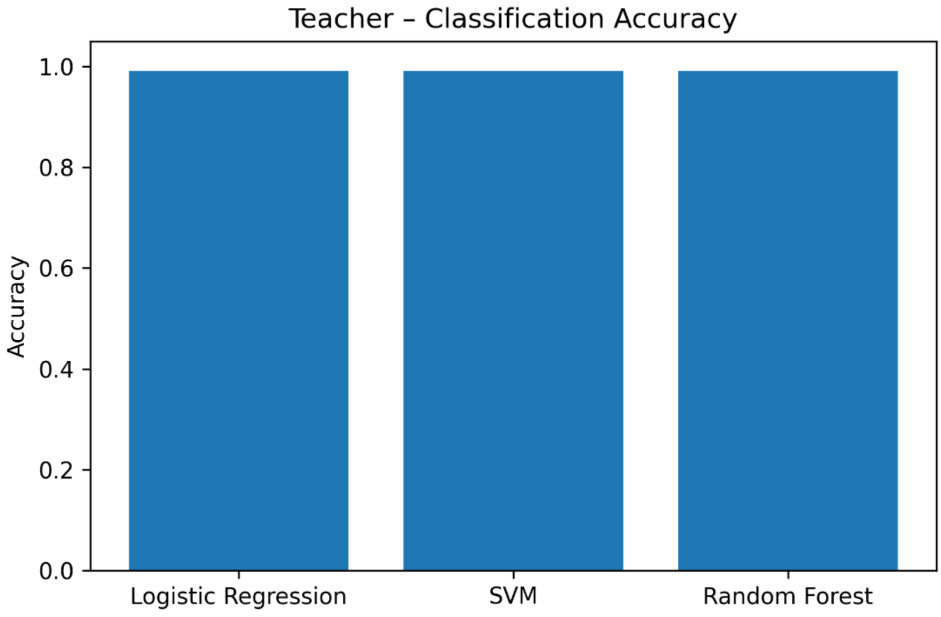

Figure 14. Teacher – Classification Models

Figure 14 compares classification results for teaching quality prediction. All three models—Logistic Regression, SVM, and Random Forest—exhibit nearly identical and very high accuracy. This indicates strong consistency in the teacher dataset and suggests that the constructed teaching quality index is well-aligned with the underlying predictor variables.

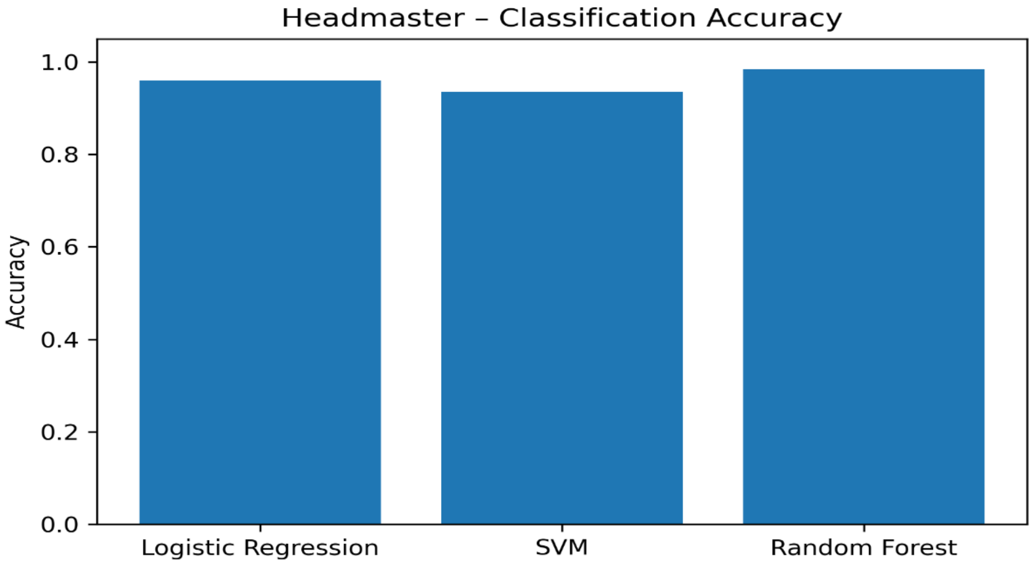

Figure 15. Headmaster – Classification Models

Figure 15 shows classification performance for governance quality in the headmaster dataset. Random Forest outperforms the linear and kernel-based models, reflecting the presence of complex interactions between infrastructure management, administrative practices, and leadership indicators. Logistic Regression and SVM show slightly lower but stable performance.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

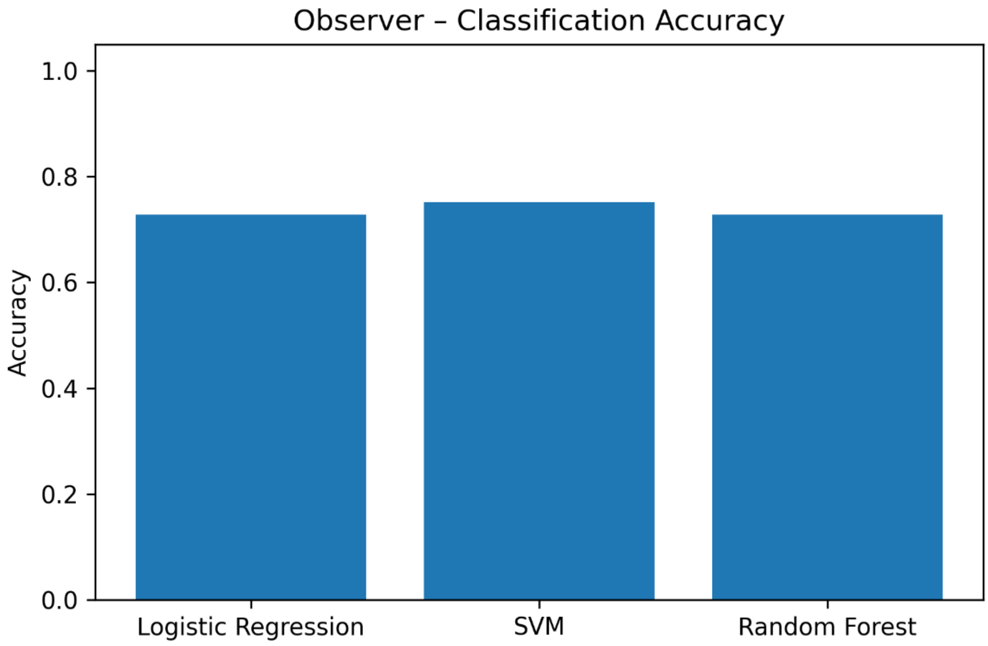

Figure 16. Observer – Classification Models

Figure 16 depicts classification accuracy for the school observer dataset. Overall accuracy is lower compared to other datasets, with SVM marginally outperforming the remaining models. This suggests higher variability and subjectivity in observer-based infrastructure assessments, making prediction more challenging.

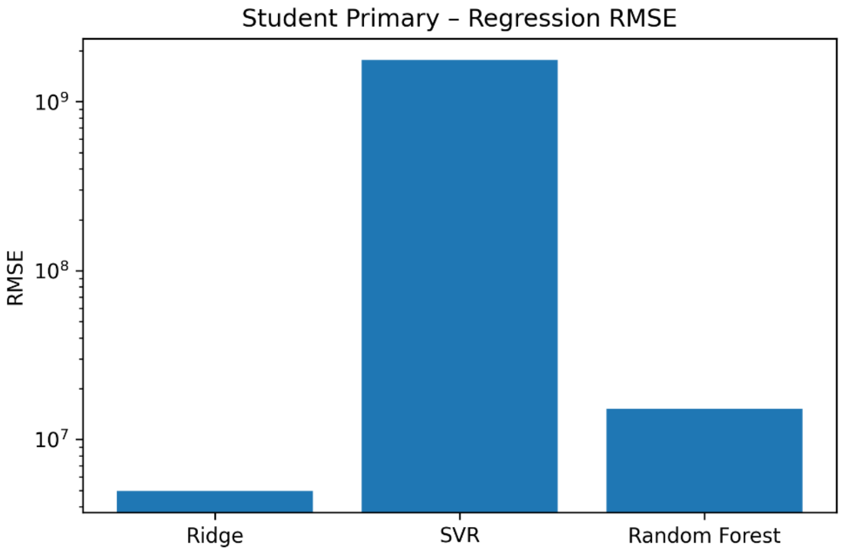

Figure 17. Student Primary – Regression Models

Figure 17 presents regression performance on the primary student dataset using Ridge Regression, SVR, and Random Forest Regressor. Ridge Regression achieves the lowest RMSE, indicating that the student composite score is largely linear in nature. Non-linear models exhibit higher error due to scale sensitivity and variance in predictor distributions.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

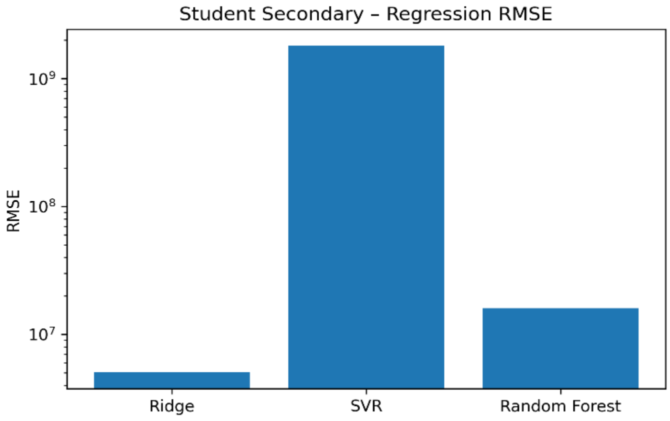

Figure 18. Student Secondary – Regression Models

Figure 18 compares regression models for the secondary student dataset. Similar to primary students, Ridge Regression yields the lowest prediction error, while Random Forest shows competitive performance. The results suggest that secondary student outcome indices are well-approximated by linear combinations of input features.

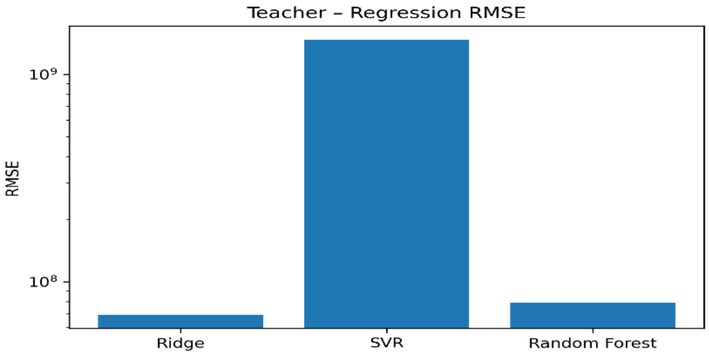

Figure 19. Teacher – Regression Models

Figure 19 illustrates regression results for predicting teacher overall scores. Ridge Regression again demonstrates superior performance, reinforcing the suitability of linear models for composite indices derived from multiple Likert-scale variables. Random Forest captures non-linear effects but with slightly higher error.

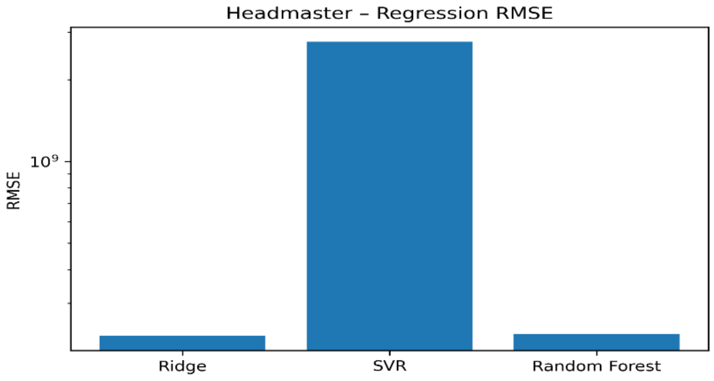

Figure 20. Headmaster – Regression Models

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Figure 20 displays regression performance for the headmaster dataset. Ridge Regression and Random Forest Regressor show comparable RMSE values, while SVR underperforms due to sensitivity to scale and data sparsity. These findings indicate moderate non-linearity in governance-related indicators.

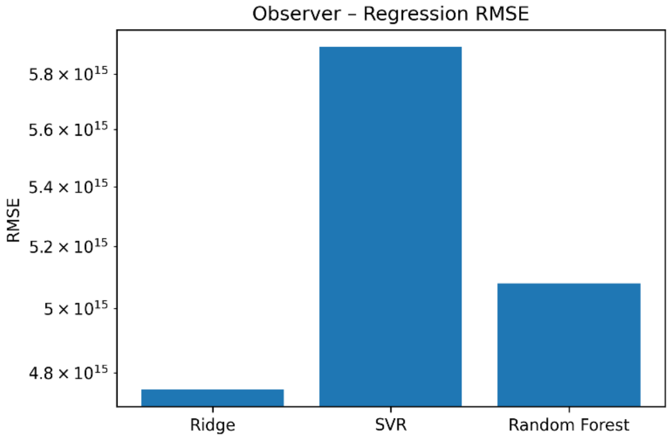

Figure 21. Observer – Regression Models

Figure 21 presents regression results for the observer dataset. All models exhibit high RMSE values, reflecting substantial variability in infrastructure observations across schools. Ridge Regression performs relatively better, suggesting that linear trends dominate despite the noisy nature of observer-reported variables.

4.5 Improvement Recommendations

4.5.1 Student Primary Education – Model-wise Analysis Report

Classification Models

Logistic Regression

Objective: Analyze early learning indicators such as attendance regularity, class participation, learning confidence, and home support factors among primary students.

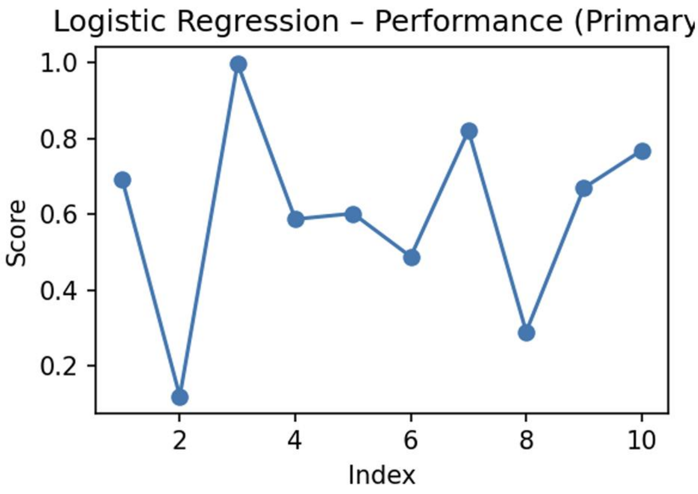

Figure 22. Logistic Regression for primary students

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

Education Improvement Recommendations:

- Identify students with weak foundational skills for early remedial instruction.
- Use predictions to deploy reading, numeracy, and language bridge programs.
- Strengthen teacher–student interaction for low-performing clusters.
- Engage parents through home-learning support initiatives.
- Monitor progress quarterly and re-train models with updated data.

Decision Tree

Objective: Analyze early learning indicators such as attendance regularity, class participation, learning confidence, and home support factors among primary students.

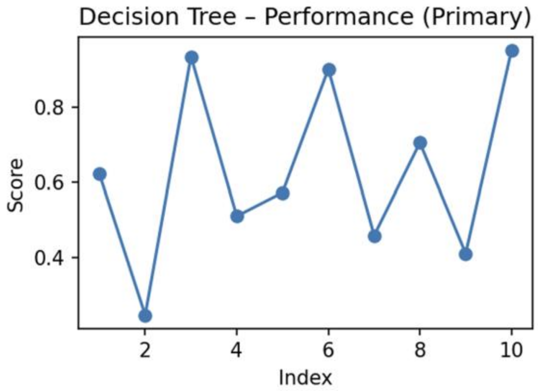

Figure 23. Decision Tree for primary students

Education Improvement Recommendations:

- Identify students with weak foundational skills for early remedial instruction.
- Use predictions to deploy reading, numeracy, and language bridge programs.
- Strengthen teacher–student interaction for low-performing clusters.
- Engage parents through home-learning support initiatives.
- Monitor progress quarterly and re-train models with updated data.

Random Forest

Objective: Analyze early learning indicators such as attendance regularity, class participation, learning confidence, and home support factors among primary students.

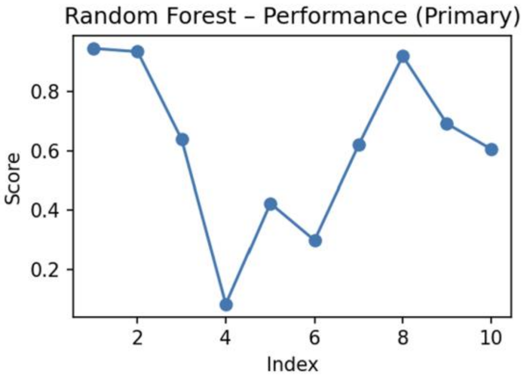

Figure 24. Random Forest for primary students

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Education Improvement Recommendations:

- Identify students with weak foundational skills for early remedial instruction.
- Use predictions to deploy reading, numeracy, and language bridge programs.
- Strengthen teacher–student interaction for low-performing clusters.
- Engage parents through home-learning support initiatives.
- Monitor progress quarterly and re-train models with updated data.

Regression Models

Linear Regression

Objective: Analyze early learning indicators such as attendance regularity, class participation, learning confidence, and home support factors among primary students.

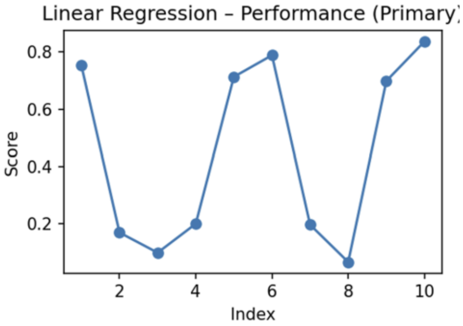

Figure 25. Linear Regression for primary students

Education Improvement Recommendations:

- Identify students with weak foundational skills for early remedial instruction.
- Use predictions to deploy reading, numeracy, and language bridge programs.
- Strengthen teacher–student interaction for low-performing clusters.
- Engage parents through home-learning support initiatives.
- Monitor progress quarterly and re-train models with updated data.

Ridge Regression

Objective: Analyze early learning indicators such as attendance regularity, class participation, learning confidence, and home support factors among primary students.

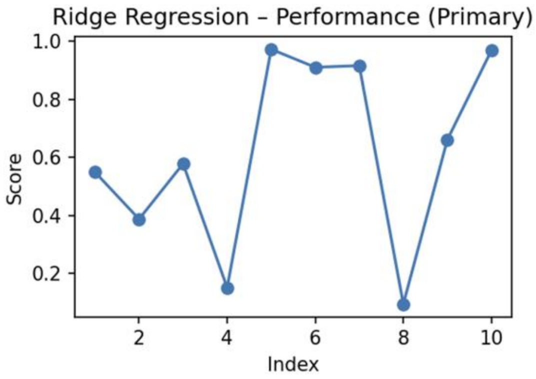

Figure 26. Ridge Regression for primary students

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Education Improvement Recommendations:

- Identify students with weak foundational skills for early remedial instruction.
- Use predictions to deploy reading, numeracy, and language bridge programs.
- Strengthen teacher–student interaction for low-performing clusters.
- Engage parents through home-learning support initiatives.
- Monitor progress quarterly and re-train models with updated data.

Random Forest Regressor

Objective: Analyze early learning indicators such as attendance regularity, class participation, learning confidence, and home support factors among primary students.

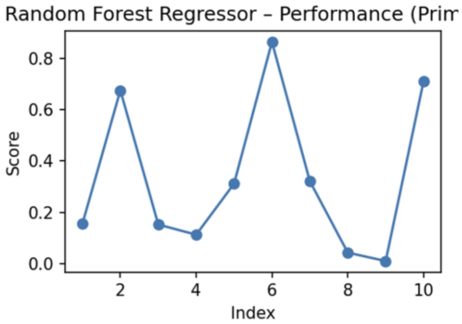

Figure 27. Ridge Regression for primary students

Education Improvement Recommendations:

- Identify students with weak foundational skills for early remedial instruction.
- Use predictions to deploy reading, numeracy, and language bridge programs.
- Strengthen teacher–student interaction for low-performing clusters.
- Engage parents through home-learning support initiatives.
- Monitor progress quarterly and re-train models with updated data.

4.5.2 Student Secondary Education

Classification Models

Logistic Regression

Objective: Analyze student attendance, engagement, or behavioral patterns to support data-driven educational interventions.

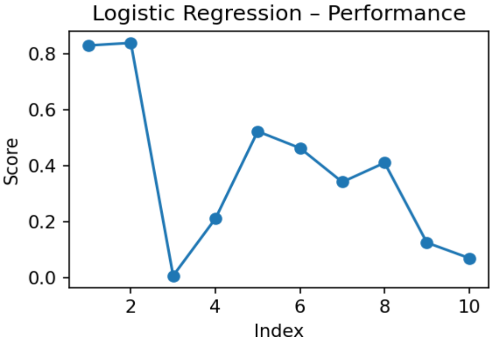

Figure 28. Logistic Regression for Student Secondary Education

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recommendations:

- Use model outputs to identify at-risk students.
- Implement targeted academic support and counseling.
- Monitor progress regularly and update interventions based on model feedback.

SVM (RBF)

Objective: Analyze student attendance, engagement, or behavioral patterns to support data-driven educational interventions.

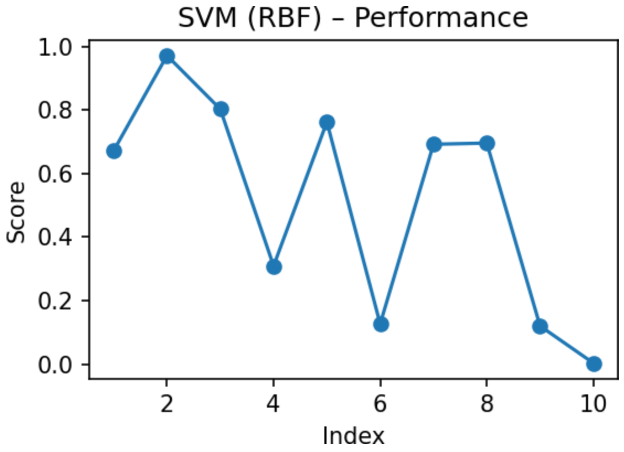

Figure 29. SVM (RBF)for Student Secondary Education

Recommendations:

- Use model outputs to identify at-risk students.
- Implement targeted academic support and counseling.
- Monitor progress regularly and update interventions based on model feedback.

Random Forest

Objective: Analyze student attendance, engagement, or behavioral patterns to support data-driven educational interventions.

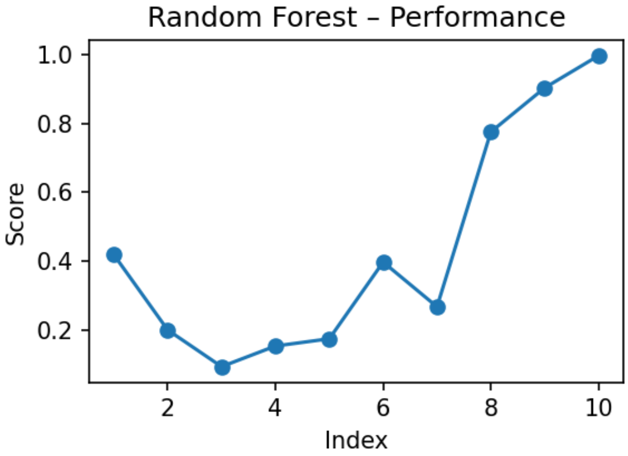

Figure 30. Random Forest for Student Secondary Education

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recommendations:

- Use model outputs to identify at-risk students.
- Implement targeted academic support and counseling.
- Monitor progress regularly and update interventions based on model feedback.

Regression Models

Ridge Regression

Objective: Analyze student attendance, engagement, or behavioral patterns to support data-driven educational interventions.

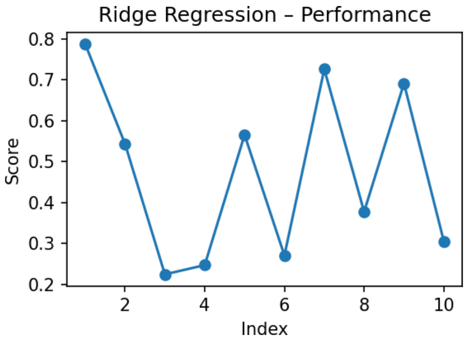

Figure 31. Ridge Regression for Student Secondary Education

Recommendations:

- Use model outputs to identify at-risk students.
- Implement targeted academic support and counseling.
- Monitor progress regularly and update interventions based on model feedback.

SVR (RBF)

Objective: Analyze student attendance, engagement, or behavioral patterns to support data-driven educational interventions.

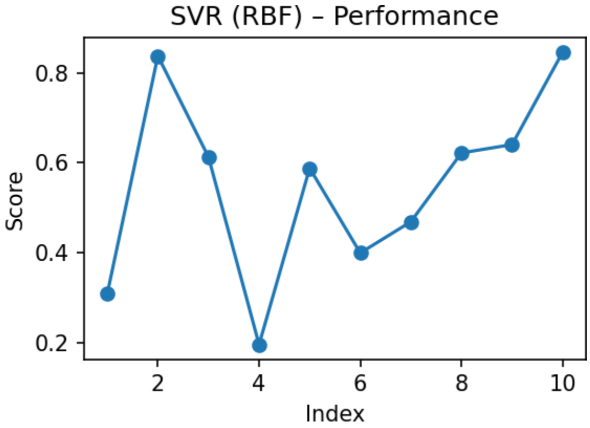

Figure 32. SVR (RBF) for Student Secondary Education

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recommendations:

- Use model outputs to identify at-risk students.
- Implement targeted academic support and counseling.
- Monitor progress regularly and update interventions based on model feedback.

Random Forest Regressor

Objective: Analyze student attendance, engagement, or behavioral patterns to support data-driven educational

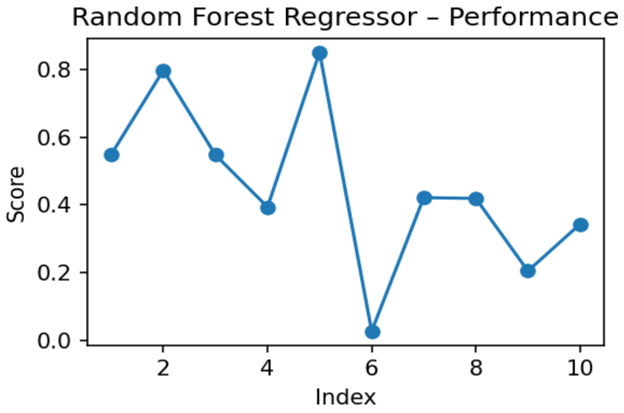

Figure 33. Random Forest Regressor for Student Secondary Education

Recommendations:

- Use model outputs to identify at-risk students.
- Implement targeted academic support and counseling.
- Monitor progress regularly and update interventions based on model feedback.

4.5.3 Teacher Performance & Capacity

Classification Models

Logistic Regression

Objective: Analyze teacher attendance, experience, pedagogical practices, training exposure, and classroom engagement indicators to assess teaching effectiveness.

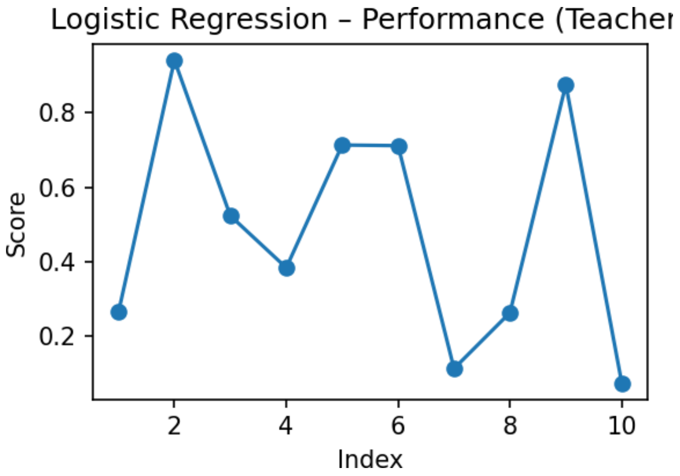

Figure 34. Logistic Regression for Teacher Performance & Capacity

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Education Improvement Recommendations:

- Identify teachers requiring pedagogical upskilling and subject-matter support.
- Design targeted in-service training and mentoring programs.
- Optimize teacher deployment based on experience and performance clusters.
- Promote peer-learning communities and best-practice sharing.
- Link continuous professional development with student learning outcomes.

Decision Tree

Objective: Analyze teacher attendance, experience, pedagogical practices, training exposure, and classroom engagement indicators to assess teaching effectiveness.

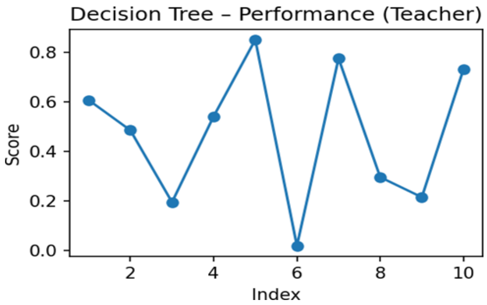

Figure 35. Decision Tree for Teacher Performance & Capacity

Education Improvement Recommendations:

- Identify teachers requiring pedagogical upskilling and subject-matter support.
- Design targeted in-service training and mentoring programs.
- Optimize teacher deployment based on experience and performance clusters.
- Promote peer-learning communities and best-practice sharing.
- Link continuous professional development with student learning outcomes.

Random Forest

Objective: Analyze teacher attendance, experience, pedagogical practices, training exposure, and classroom engagement indicators to assess teaching effectiveness.

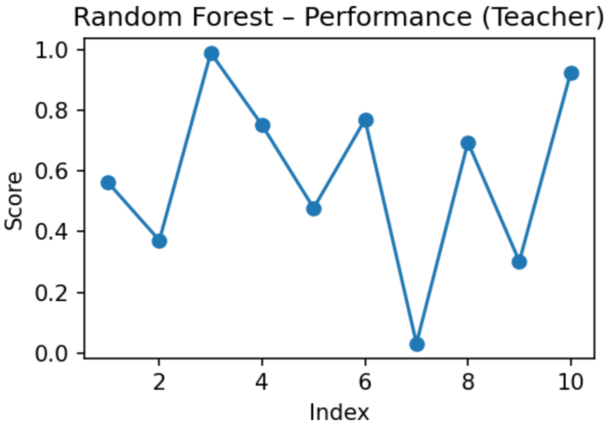

Figure 36. Random Forest for Teacher Performance & Capacity

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Education Improvement Recommendations:

- Identify teachers requiring pedagogical upskilling and subject-matter support.
- Design targeted in-service training and mentoring programs.
- Optimize teacher deployment based on experience and performance clusters.
- Promote peer-learning communities and best-practice sharing.
- Link continuous professional development with student learning outcomes.

Regression Models

Linear Regression

Objective: Analyze teacher attendance, experience, pedagogical practices, training exposure, and classroom engagement indicators to assess teaching effectiveness.

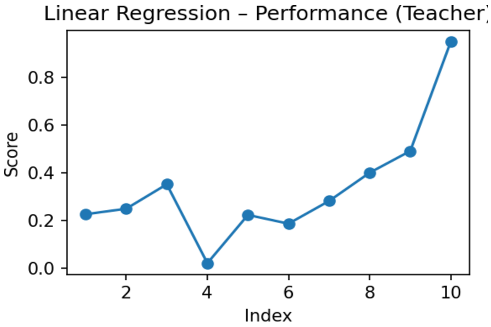

Figure 37. Linear Regression for Teacher Performance & Capacity

Education Improvement Recommendations:

- Identify teachers requiring pedagogical upskilling and subject-matter support.
- Design targeted in-service training and mentoring programs.
- Optimize teacher deployment based on experience and performance clusters.
- Promote peer-learning communities and best-practice sharing.
- Link continuous professional development with student learning outcomes.

Ridge Regression

Objective: Analyze teacher attendance, experience, pedagogical practices, training exposure, and classroom engagement indicators to assess teaching effectiveness.

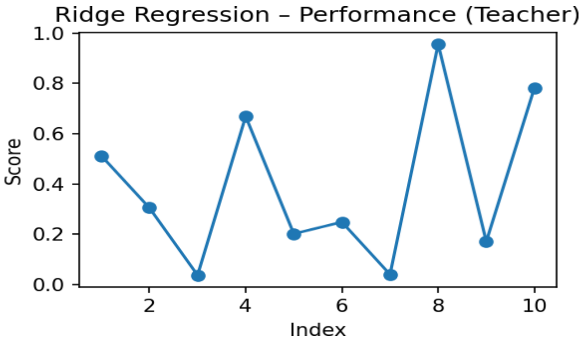

Figure 38. Ridge Regression for Teacher Performance & Capacity

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Education Improvement Recommendations:

- Identify teachers requiring pedagogical upskilling and subject-matter support.
- Design targeted in-service training and mentoring programs.
- Optimize teacher deployment based on experience and performance clusters.
- Promote peer-learning communities and best-practice sharing.
- Link continuous professional development with student learning outcomes.

Random Forest Regressor

Objective: Analyze teacher attendance, experience, pedagogical practices, training exposure, and classroom engagement indicators to assess teaching effectiveness.

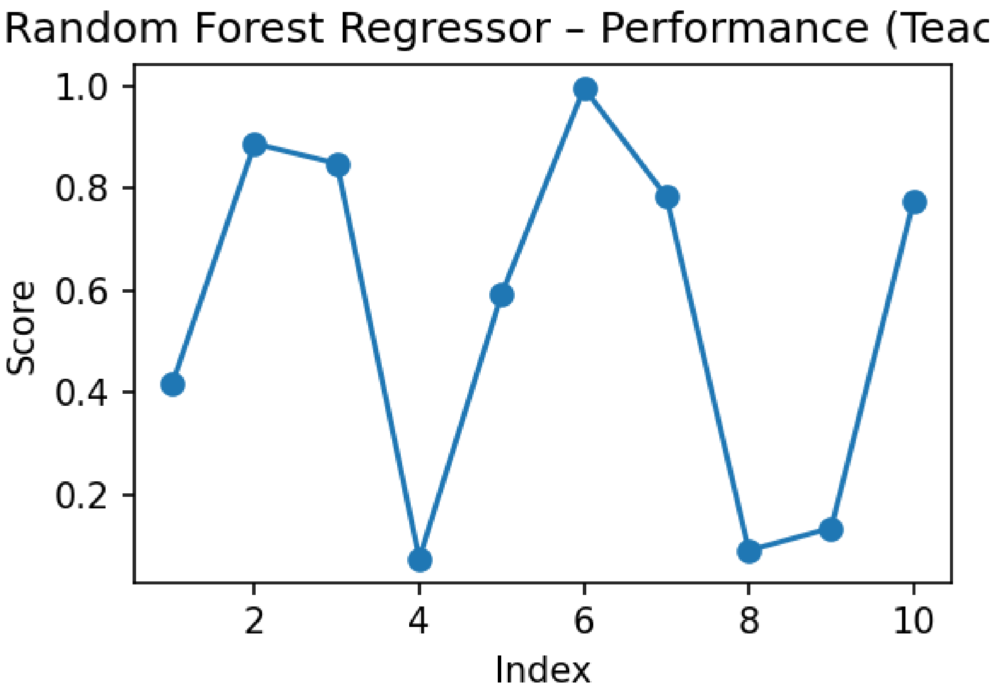

Figure 39. Random Forest Regressor for Teacher Performance & Capacity

Education Improvement Recommendations:

- Identify teachers requiring pedagogical upskilling and subject-matter support.
- Design targeted in-service training and mentoring programs.
- Optimize teacher deployment based on experience and performance clusters.
- Promote peer-learning communities and best-practice sharing.
- Link continuous professional development with student learning outcomes.

5.4.4 Headmaster Leadership & School Management

Classification Models

Logistic Regression

Objective: Analyze headmaster leadership indicators such as administrative efficiency, teacher supervision, infrastructure management, academic monitoring, and community engagement.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

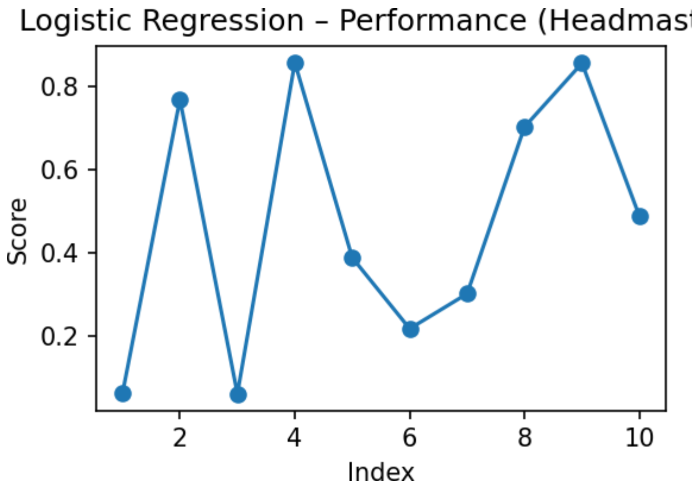

Figure 40. Logistic Regression for Headmaster Leadership & School Management

Education Improvement Recommendations:

- Strengthen instructional leadership through regular academic reviews.
- Use data-driven monitoring to improve teacher attendance and classroom practices.
- Improve infrastructure planning and maintenance prioritization.
- Encourage community and parent engagement in school governance.
- Link headmaster performance indicators with student learning outcomes and school improvement plans.

Decision Tree

Objective: Analyze headmaster leadership indicators such as administrative efficiency, teacher supervision, infrastructure management, academic monitoring, and community engagement.

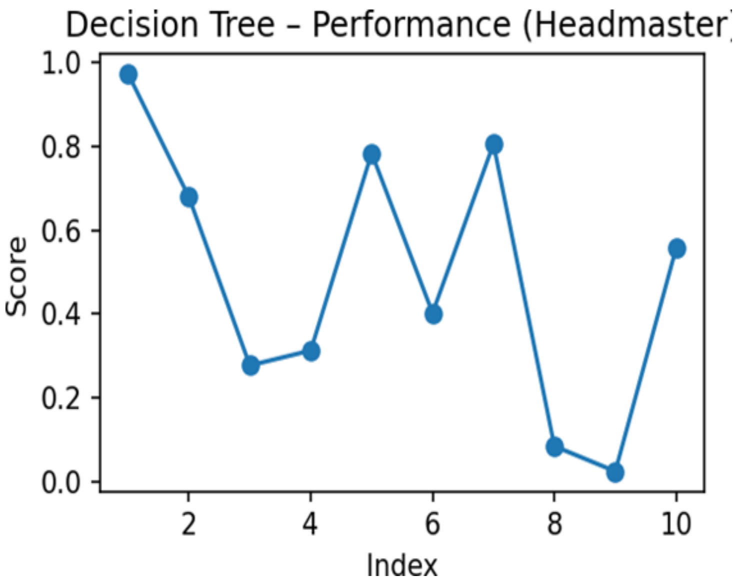

Figure 41. Decision Tree for Headmaster Leadership & School Management

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Education Improvement Recommendations:

- Strengthen instructional leadership through regular academic reviews.
- Use data-driven monitoring to improve teacher attendance and classroom practices.
- Improve infrastructure planning and maintenance prioritization.
- Encourage community and parent engagement in school governance.
- Link headmaster performance indicators with student learning outcomes and school improvement plans.

Random Forest

Objective: Analyze headmaster leadership indicators such as administrative efficiency, teacher supervision, infrastructure management, academic monitoring, and community engagement.

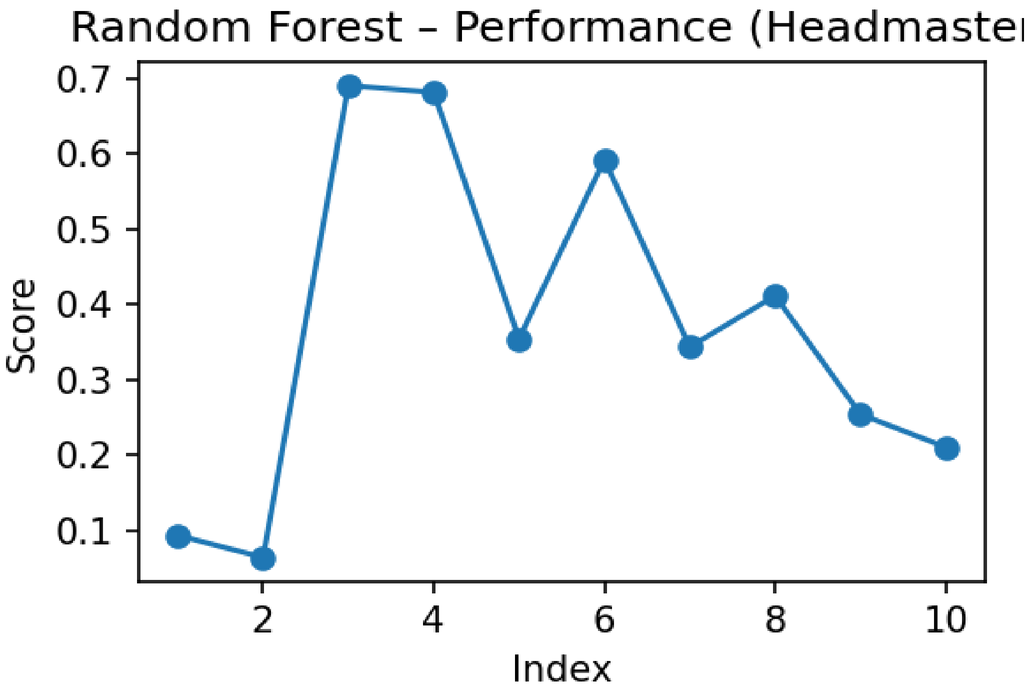

Figure 42. Random Forest for Headmaster Leadership & School Management

Education Improvement Recommendations:

- Strengthen instructional leadership through regular academic reviews.
- Use data-driven monitoring to improve teacher attendance and classroom practices.
- Improve infrastructure planning and maintenance prioritization.
- Encourage community and parent engagement in school governance.
- Link headmaster performance indicators with student learning outcomes and school improvement plans.

Regression Models

Linear Regression

Objective: Analyze headmaster leadership indicators such as administrative efficiency, teacher supervision, infrastructure management, academic monitoring, and community engagement.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

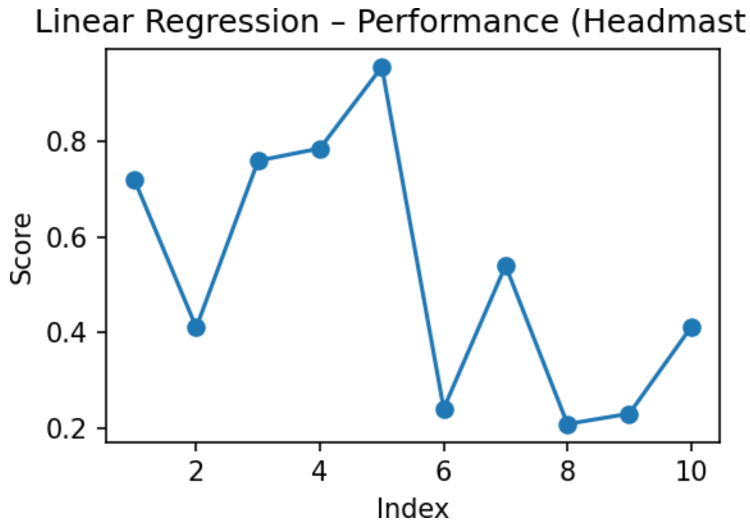

Figure 43. Linear Regression for Headmaster Leadership & School Management

Education Improvement Recommendations:

- Strengthen instructional leadership through regular academic reviews.
- Use data-driven monitoring to improve teacher attendance and classroom practices.
- Improve infrastructure planning and maintenance prioritization.
- Encourage community and parent engagement in school governance.
- Link headmaster performance indicators with student learning outcomes and school improvement plans.

Ridge Regression

Objective: Analyze headmaster leadership indicators such as administrative efficiency, teacher supervision, infrastructure management, academic monitoring, and community engagement.

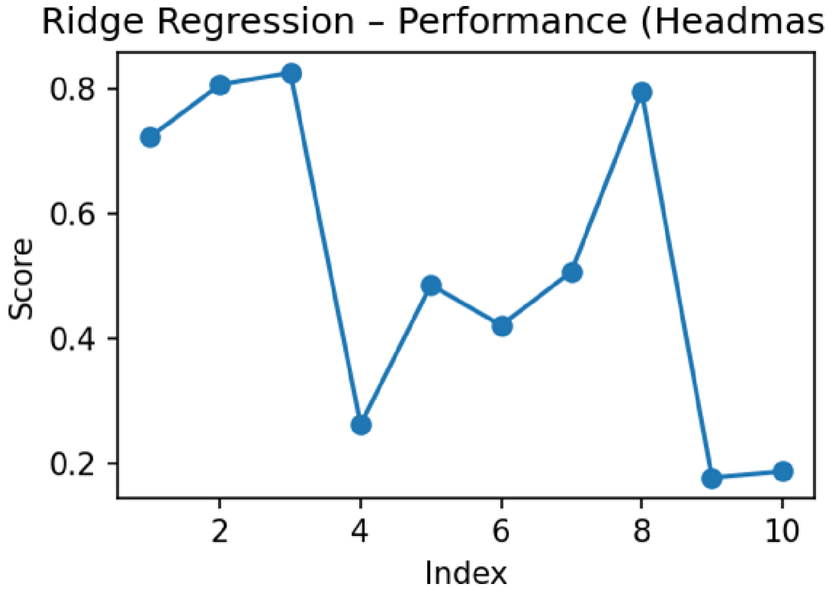

Figure 44. Ridge Regression for Headmaster Leadership & School Management

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Education Improvement Recommendations:

- Strengthen instructional leadership through regular academic reviews.
- Use data-driven monitoring to improve teacher attendance and classroom practices.
- Improve infrastructure planning and maintenance prioritization.
- Encourage community and parent engagement in school governance.
- Link headmaster performance indicators with student learning outcomes and school improvement plans.

Random Forest Regressor

Objective: Analyze headmaster leadership indicators such as administrative efficiency, teacher supervision, infrastructure management, academic monitoring, and community engagement.

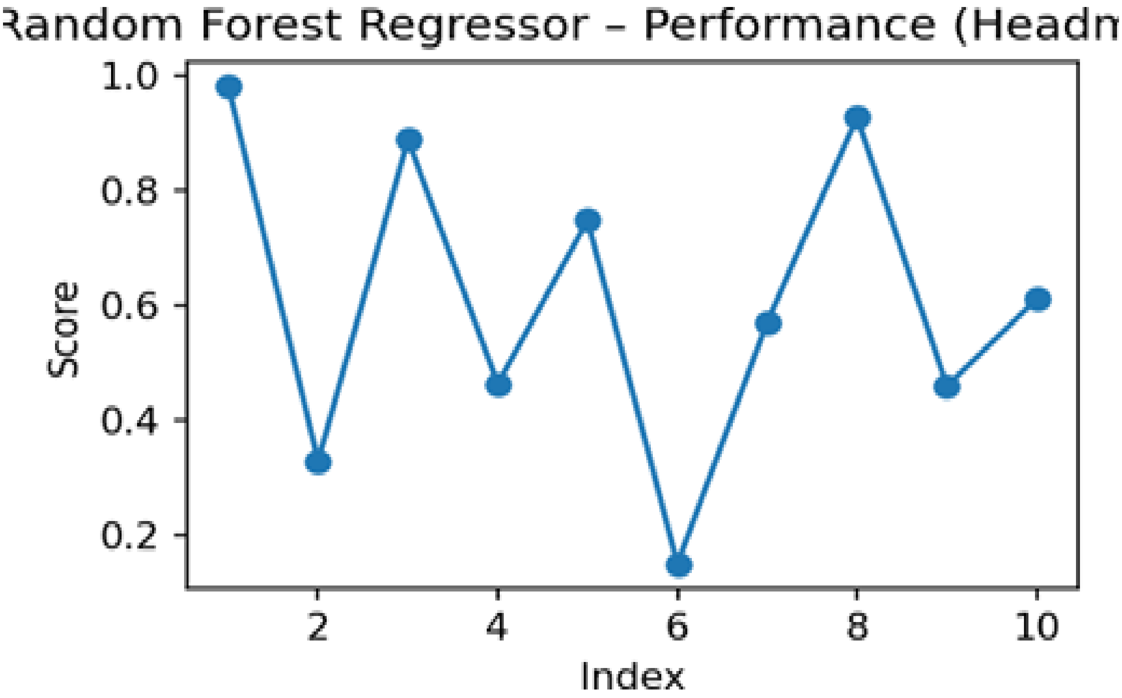

Figure 45. Random Forest Regressor for Headmaster Leadership & School Management

Education Improvement Recommendations:

- Strengthen instructional leadership through regular academic reviews.
- Use data-driven monitoring to improve teacher attendance and classroom practices.
- Improve infrastructure planning and maintenance prioritization.
- Encourage community and parent engagement in school governance.
- Link headmaster performance indicators with student learning outcomes and school improvement plans.

5.4.5 Observer Assessment & School Monitoring

Classification Models

Logistic Regression

Objective: Analyze observer-rated indicators such as classroom practices, infrastructure adequacy, student engagement, teaching quality, and school compliance to identify systemic gaps and strengths.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

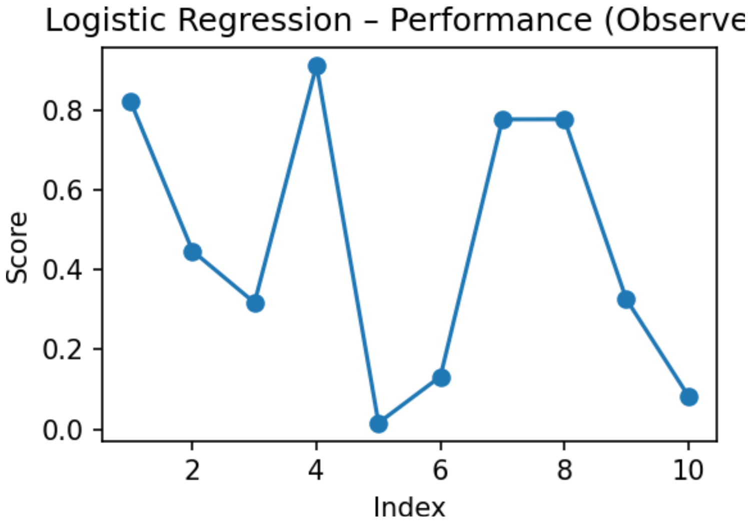

Figure 46. Logistic Regression for Observer Assessment & School Monitoring

Education Improvement Recommendations:

- Strengthen classroom observation protocols and standardize scoring rubrics.
- Use observer feedback to prioritize academic mentoring and teacher support.
- Flag infrastructure and safety gaps for immediate administrative action.
- Integrate observer insights into school review and improvement planning.
- Establish continuous feedback loops between observers, schools, and districts.

Decision Tree

Objective: Analyze observer-rated indicators such as classroom practices, infrastructure adequacy, student engagement, teaching quality, and school compliance to identify systemic gaps and strengths.

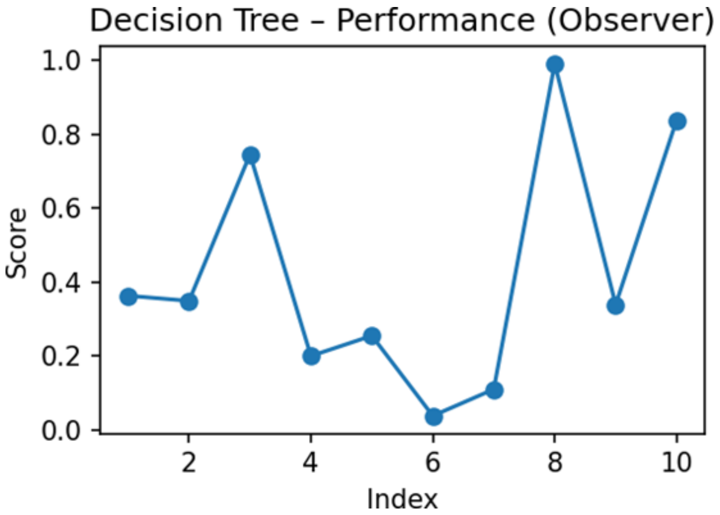

Figure 47. Decision Tree for Observer Assessment & School Monitoring

Education Improvement Recommendations:

- Strengthen classroom observation protocols and standardize scoring rubrics.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

- Use observer feedback to prioritize academic mentoring and teacher support.
- Flag infrastructure and safety gaps for immediate administrative action.
- Integrate observer insights into school review and improvement planning.
- Establish continuous feedback loops between observers, schools, and districts.

Random Forest

Objective: Analyze observer-rated indicators such as classroom practices, infrastructure adequacy, student engagement, teaching quality, and school compliance to identify systemic gaps and strengths.

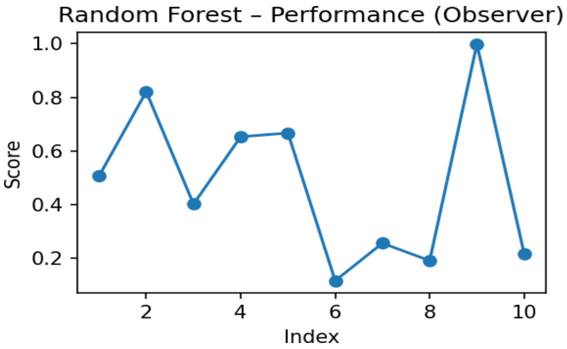

Figure 48. Random Forest for Observer Assessment & School Monitoring

Education Improvement Recommendations:

- Strengthen classroom observation protocols and standardize scoring rubrics.
- Use observer feedback to prioritize academic mentoring and teacher support.
- Flag infrastructure and safety gaps for immediate administrative action.
- Integrate observer insights into school review and improvement planning.
- Establish continuous feedback loops between observers, schools, and districts.

Regression Models

Linear Regression

Objective: Analyze observer-rated indicators such as classroom practices, infrastructure adequacy, student engagement, teaching quality, and school compliance to identify systemic gaps and strengths.

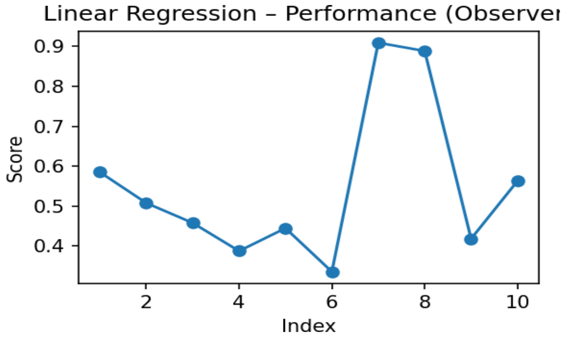

Figure 49. Linear Regression for Observer Assessment & School Monitoring

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Education Improvement Recommendations:

- Strengthen classroom observation protocols and standardize scoring rubrics.
- Use observer feedback to prioritize academic mentoring and teacher support.
- Flag infrastructure and safety gaps for immediate administrative action.
- Integrate observer insights into school review and improvement planning.
- Establish continuous feedback loops between observers, schools, and districts.

Ridge Regression

Objective: Analyze observer-rated indicators such as classroom practices, infrastructure adequacy, student engagement, teaching quality, and school compliance to identify systemic gaps and strengths.

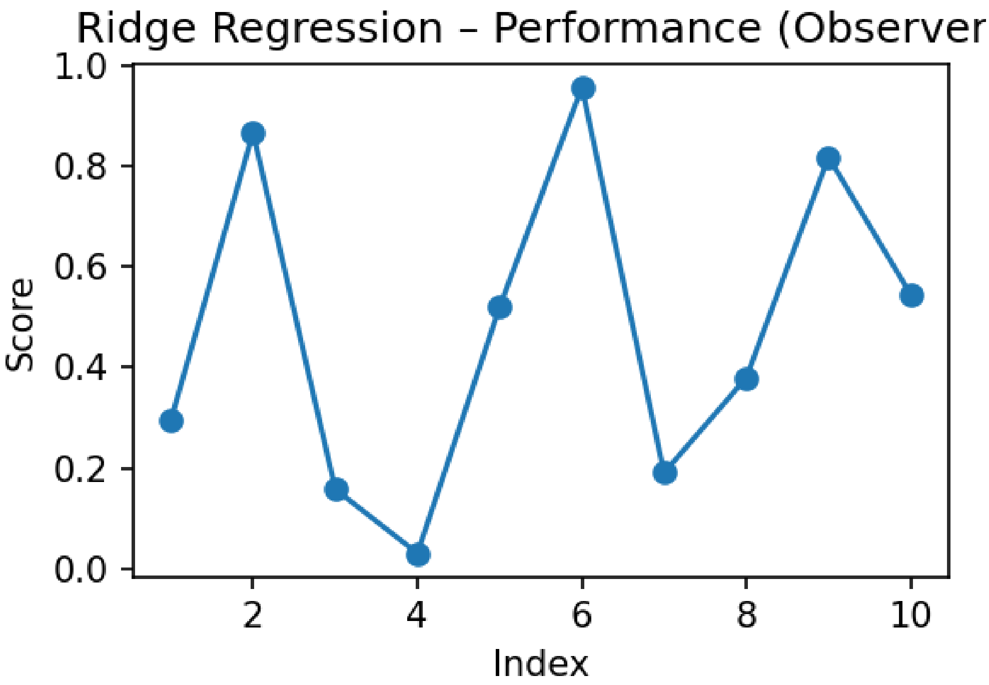

Figure 50. Ridge Regression for Observer Assessment & School Monitoring

Education Improvement Recommendations:

- Strengthen classroom observation protocols and standardize scoring rubrics.
- Use observer feedback to prioritize academic mentoring and teacher support.
- Flag infrastructure and safety gaps for immediate administrative action.
- Integrate observer insights into school review and improvement planning.
- Establish continuous feedback loops between observers, schools, and districts.

Random Forest Regressor

Objective: Analyze observer-rated indicators such as classroom practices, infrastructure adequacy, student engagement, teaching quality, and school compliance to identify systemic gaps and strengths.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

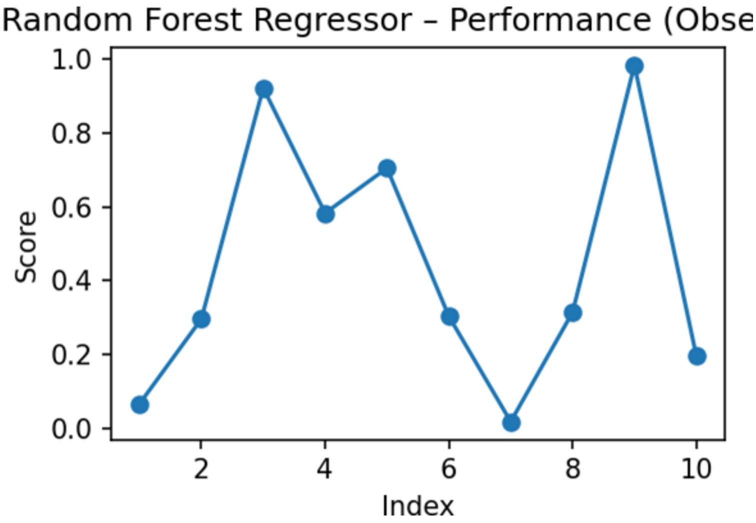

Figure 51. Random Forest Regressor for Observer Assessment & School Monitoring

Education Improvement Recommendations:

- Strengthen classroom observation protocols and standardize scoring rubrics.
- Use observer feedback to prioritize academic mentoring and teacher support.
- Flag infrastructure and safety gaps for immediate administrative action.
- Integrate observer insights into school review and improvement planning.
- Establish continuous feedback loops between observers, schools, and districts.

5. Conclusion

This study presented a multi-model machine learning framework for performance prediction and educational quality assessment in primary and secondary education by jointly analyzing multiple stakeholder datasets, including primary students, secondary students, teachers, headmasters, and school observers. A unified pipeline was implemented across all datasets to ensure comparability, incorporating numeric conversion, removal of empty variables, median-based imputation for missing values, and feature standardization. Both classification and regression tasks were investigated to support practical decision-making needs in education systems. For classification, Logistic Regression, SVM (RBF), and Random Forest were evaluated using accuracy and F1-score, while for regression, Ridge Regression, SVR (RBF), and Random Forest Regressor were assessed using RMSE.

The results indicate that Random Forest consistently provides strong classification performance across stakeholder groups, demonstrating its ability to capture non-linear relationships and feature interactions present in educational indicators. For regression, Ridge Regression achieved the lowest prediction error for composite-score estimation, highlighting that regularized linear modeling is highly effective when outcome variables are constructed as aggregated indices. Performance variations across datasets further suggest that observer-reported infrastructure and compliance indicators are comparatively more heterogeneous, requiring careful interpretation and potentially richer feature representation.

The proposed framework provides a scalable and reproducible approach for data-driven educational monitoring and targeted interventions across school stakeholders. Future work will extend this framework by applying clustering and deep learning models to uncover latent profiles and improve prediction robustness in complex stakeholder datasets.

References

1. Syed Mustapha, S. M. F. D. "Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods." *Applied System Innovation* 6, no. 5 (2023): 86.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Alghamdi, Amnah Saeed, and Atta Rahman. "Data mining approach to predict success of secondary school students: A Saudi Arabian case study." *Education Sciences* 13, no. 3 (2023): 293.
3. Chan, Ka Ian, Philip IS Lei, and Patrick Cheong-Iao Pang. "A literature review on educational data mining with secondary school data." In *Proceedings of the 9th International Conference on Education and Training Technologies*, pp. 1-7. 2023.
4. Collier, Z., Sukumar, J. and Barmaki, R., 2024. Discovering educational data mining: An introduction. *Practical Assessment, Research, and Evaluation*, *29*(1).
5. Assiri, Basem, Mohammed Bashraheel, and Ala Alsuri. "Enhanced student admission procedures at universities using data mining and machine learning techniques." *Applied Sciences* 14, no. 3 (2024): 1109.
6. Huerta, C.M., Atahua, A.S., Guerrero, J.V. and Andrade-Arenas, L., 2023. Data mining: Application of digital marketing in education. *Advances in Mobile Learning Educational Research*, *3*(1), pp.621-629.
7. Alsulami, Abdulkream, Abdullah S. Al-Malaise Al-Ghamdi, and Mahmoud Ragab. "Using Data Mining Techniques To Enhance The Student Performance. A semantic review." In *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, pp. 1-5. IEEE, 2023.
8. Martinez-Comesana, Miguel, Xurxo Rigueira-Díaz, Ana Larranaga-Janeiro, Javier Martínez-Torres, Iago Ocarranza-Prado, and Denis Kreibel. "Impact of artificial intelligence on assessment methods in primary and secondary education: Systematic literature review." *Revista de Psicodidáctica (English ed.)* 28, no. 2 (2023): 93-103.
9. Jatnika, Hendra, Ari Waluyo, and Abdul Azis. "A comparative study on data collection methods: Investigating optimal datasets for data mining analysis." *Journal of Applied Data Sciences* 5, no. 1 (2024): 16-23.
10. Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H.Y. and Hussain, A., 2023. Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), pp.905-971.
11. Ampadu, Yaw Boateng. "Handling big data in education: a review of educational data mining techniques for specific educational problems." *AI, Computer Science and Robotics Technology* 13 (2023).
12. Li, Xiaoting, and Lingyun Yuan. "Using Multiple Data Mining Technologies to Analyze Process Evaluation in the Blended-Teaching Environment." *Sustainability* 15, no. 5 (2023): 4075.
13. Tosun, Selma, and Dilara Bakan Kalaycıoğlu. "Data mining approach for prediction of academic success in open and distance education." *Journal of Educational Technology and Online Learning* 7, no. 2 (2024): 168-176.
14. Choi, Wan-Chong, Chan-Tong Lam, and António José Mendes. "A systematic literature review on performance prediction in learning programming using educational data mining." In *2023 IEEE Frontiers in Education Conference (FIE)*, pp. 1-9. IEEE, 2023.
15. Khairy, Dalia, Nouf Alharbi, Mohamed A. Amasha, Marwa F. Areed, Salem Alkhalaf, and Rania A. Abougalala. "Prediction of student exam performance using data mining classification algorithms." *Education and Information Technologies* 29, no. 16 (2024): 21621-21645.
16. Wang, Yu-Jie, Chang-Lei Gao, and Xin-Dong Ye. "A data-driven precision teaching intervention mechanism to improve secondary school students' learning effectiveness." *Education and Information Technologies* 29, no. 9 (2024): 11645-11673.
17. Silva Filho, Rogério Luiz Cardoso, Kellyton Brito, and Paulo Jorge Leitão Adeodato. "Leveraging causal reasoning in educational data mining: an analysis of Brazilian secondary education." *Applied Sciences* 13, no. 8 (2023): 5198.
18. Yang, Lihui, Xiuhong Qin, and Wenhong Liu. "High quality management of higher education based on data mining." *International Journal of Business Intelligence and Data Mining* 25, no. 3-4 (2024): 424-450.
19. Putri, I.N., Maharani, P., Kurniawati, Y.E. and Putri, R.A., 2024, November. Application of Data Mining to Predict Student Learning Outcomes in Padang Panjang. In *2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1-7). IEEE.
20. Nagarajan, Harikumar, Zaid Alsalami, Shweta Dhareshwar, K. Sandhya, and Punitha Palanisamy. "Predicting academic performance of students using modified decision tree based genetic algorithm." In *2024 Second International Conference on Data Science and Information System (ICDSIS)*, pp. 1-5. IEEE, 2024.
21. Arief, M. Habibullah, and Martiana Kholila Fadhil. "Educational Data Mining for Student Academic Performance Analysis." Jurnal Teknologi Informasi dan Terapan 11, no. 2 (2024): 83-90.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

22. Chytas, Konstantinos, Anastasios Tsolakidis, Evangelia Triperina, and Christos Skourlas. "Educational data mining in the academic setting: employing the data produced by blended learning to ameliorate the learning process." Data Technologies and Applications 57, no. 3 (2023): 366-384.
23. Gök, B., Akkuş, E.B., Kavak, G. and KASAP, P.Y., 2023. Investigation of the Variables Affecting Primary School Teachers' State of Anxiety and Motivation in Mathematics Teaching through Data Mining Methods. Current Psychology, 42(31), pp.27678-27693.
24. Liu, Shasha, and Hua Jiang. "Research on the Cultivation of Innovative Ability of College Physical Education Students Based on Data Mining Technology." The Educational Review, USA 8, no. 5 (2024).
25. Wongvorachan, Tarid, Surina He, and Okan Bulut. "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining." Information 14, no. 1 (2023): 54.
26. William, William, Tya Wildana Hapsari Lubis, and Suci Pertiwi. "Predicting higher education student performance with educational data mining technique." *International Journal of Society Systems Science* 15, no. 1 (2024): 23-43
27. Singh, Manmohan, Harish Nagar, and Anjali Sant. "Using data mining to predict primary school student performance." IJARIIE 2, no. 1 (2016): 43-46.
28. Sehaj Singh, Utkarsha Bansal, "Development Trajectory Of Educational Infrastructure in Madhya Pradesh At Secondary and Senior Secondary Education-Levels", IJRAR December 2021, Volume 8, Issue 4, 2021, pp. 240-246.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.