Federated and Cross-Domain Student Performance Prediction

Sam Zhe Xuan[1], P. Ganesh Kumar[2], C. Rani[3], Kanagalakshmi[4], R. RajiniGanth[5], Atif Mahmood[6]*
Faculty of Data Science and Information Technology, INTI International University, Malaysia[1,6]
Department of Computer Science and Engineering-College of Engineering, Guindy,
Anna University, Chennai, Tamil Nadu, India[2]
Department of Computer Science and Engineering, Government College of Engineering,
Bodinayakanur, Tamil Nadu, India[3]
School of Science and Computer Studies, CMR University, Bangalore, Karnataka, India[4]
Department of Computer Science and Technology, SNS College of Engineering, Coimbatore, Tamil Nadu, India[5]

***Abstract*—Accurate student performance prediction is critical for data-driven educational decision-making; however, it is often hindered by data heterogeneity, privacy constraints, and domain shift across academic contexts. This study investigates student final grade (G3) prediction using three complementary machine learning paradigms: centralized learning, cross-domain generalization, and federated learning with personalization. Experiments were conducted on Portuguese and Mathematics student datasets using traditional regression models, ensemble methods, neural networks, and a personalized federated learning framework based on FedProx and FedBN. In the centralized setting, models capable of capturing non-linear relationships, particularly XGBoost and multi-layer perceptrons, achieved superior predictive performance, with XGBoost attaining an R^2 of 0.8308 and the lowest error metrics. In contrast, direct cross-domain application of models trained on Portuguese data to Mathematics outcomes resulted in severe performance degradation, with several models yielding negative R^2 values, highlighting the adverse impact of domain shift. To address privacy and heterogeneity challenges, a federated learning simulation was implemented. While the global federated model achieved moderate accuracy, the introduction of local personalization led to substantial performance gains. The personalized client models achieved stronger local predictive performance than the global federated model and showed competitive performance relative to centralized baselines. Learning-curve analysis further indicate that model performance in centralized settings improves with increasing data size but eventually plateaus, whereas cross-domain learning remains constrained despite additional data. In federated learning, predictive performance consistently improves across training rounds, demonstrating the effectiveness of iterative collaboration and client-level personalization. Overall, the results suggest that federated learning with personalization offers a competitive privacy-preserving alternative to centralized modeling and provides a clear improvement over direct cross-domain transfer in heterogeneous educational analytics.**

***Keywords*—Federated learning; cross-domain generalization; educational data mining; student performance; quality education**

I. Introduction

Educational institutions increasingly rely on data-driven approaches to understand, predict, and improve student academic performance [1]. Accurate early prediction of student outcomes enables timely interventions, optimized resource allocation, and informed pedagogical decision-making. With the growing availability of educational datasets, machine learning (ML) techniques have become central to learning analytics, offering superior predictive capabilities compared to traditional statistical methods [2]. However, deploying such models in real educational environments introduces challenges related to data heterogeneity, privacy preservation, and model interpretability [3], [4]. The prediction of student performance is inherently complex due to the multidimensional nature of educational data, which typically includes demographic, socio-economic, behavioral, and academic attributes [5].

Prior studies have extensively investigated a range of supervised learning models for predicting student academic performance, including linear regression variants, decision trees, ensemble methods, and neural networks, reporting varying levels of predictive accuracy [6]. Empirical evidence suggests that machine learning models can achieve very high performance, with some approaches attaining predictive accuracy of up to 97.12% [7]. In particular, tree-based ensemble techniques such as Random Forests, Gradient Boosting [8], and XGBoost [7] have consistently demonstrated superior performance due to their ability to capture complex nonlinear relationships and feature interactions [9]. However, despite their strong predictive capability, accuracy alone is insufficient in educational contexts, where transparency and explainability are essential to foster trust, support informed decision-making, and encourage adoption by educators and policymakers.

In addition to interpretability concerns, data privacy has emerged as a major barrier to the large-scale deployment of predictive analytics in education. Student data are highly sensitive and often governed by strict institutional and legal constraints, limiting the feasibility of centralized data collection across schools, campuses, or educational authorities. As a result, many existing studies assume centralized access to student records, which may not reflect realistic deployment scenarios. This gap motivates the exploration of privacy-preserving learning paradigms that can leverage distributed data while respecting institutional boundaries.

Federated learning (FL) is a decentralized learning framework in which multiple clients collaboratively train a shared global model without exchanging raw data. Instead, only

*Corresponding author

model updates, such as parameters or gradients, are communicated to a central server for aggregation, ensuring data privacy. This approach is well-suited to educational analytics, where institutions often exhibit heterogeneous (non-IID) data distributions due to differences in curricula, socio-economic factors, and assessment practices. However, standard FL methods such as Federated Averaging (FedAvg) often degrade under strong data heterogeneity, motivating the use of more robust variants such as Federated Proximal (FedProx) and Federated Batch Normalization (FedBN). Recent studies have also shown the feasibility of applying federated learning in privacy-preserving and resource-constrained distributed environments, including healthcare and network-based deployment scenarios [10].

In this study, heterogeneity refers to distributional differences between the Portuguese and Mathematics student datasets, including differences in feature distributions, target distributions, and learned model behavior across domains. Domain shift is operationally assessed through the degradation of R^2, MAE, and RMSE when a model trained on Portuguese data is evaluated on Mathematics data. Privacy is addressed at the simulation level by assuming that raw client data remain local during federated training, while only model parameters or updates are shared. Therefore, the privacy claim is limited to data locality and does not imply formal privacy guarantees such as differential privacy or cryptographic protection.

A longstanding challenge in educational machine learning is balancing predictive accuracy with interpretability. Advanced models such as deep neural networks [11] and ensemble methods, including XGBoost, often achieve strong predictive performance but operate as black-box systems, limiting transparency and interpretability [12], [13]. In contrast, simpler models such as Linear Regression and Elastic Net offer clear coefficient-based interpretability but may struggle to capture the complex non-linear relationships present in educational and behavioral data [14], [15]. Building on prior studies that emphasize interpretable educational prediction frameworks [16], [17], this work extends traditional regression-based approaches to cross-domain and federated learning paradigms. Beyond centralized evaluations of Elastic Net Regression, Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), and XGBoost, two additional experimental settings are introduced to enhance generalization, fairness, and privacy across heterogeneous educational domains.

Interpretability is considered as an important contextual requirement for educational prediction systems. However, the experimental focus of this study is on comparing centralized learning, cross-domain generalization, and personalized federated learning rather than conducting a separate explainability analysis.

The main contributions of this study are as follows:

- A comprehensive comparison of regression, kernel-based, neural, and ensemble models for student performance prediction.
- Introduction of cross-domain generalization and federated learning (FedProx + FedBN) frameworks to enhance scalability, fairness, and privacy under decentralized educational settings.
- Cross-Domain Learning (POR $\rightarrow$ MAT): This experiment evaluates model generalizability by training on one subject domain (Portuguese language) and testing on another (Mathematics) within the UCI Student Performance dataset [18]. It simulates realistic academic scenarios where predictive models must adapt to diverse cohorts without retraining from scratch.
- Federated Learning (FedProx + FedBN): To address data privacy and institutional heterogeneity, a hybrid federated learning framework was implemented. The model integrates *FedProx* regularization [19] to stabilize client updates under heterogeneous federated optimization and *FedBN* [20] to preserve client-specific batch-normalization statistics under non-IID feature distributions. This approach enables decentralized training across multiple clients while supporting local personalization and data locality.

II. Related Work

Student performance prediction has been a central topic in educational data mining (EDM), with extensive studies employing machine learning (ML) algorithms to model academic outcomes. Early foundational work by Cortez and Silva [16] applied decision trees, random forests, neural networks, and regression models to the UCI Student Performance dataset, showing that prior academic grades were among the strongest predictors of final performance. The UCI repository also identifies the Cortez and Silva study as the introductory paper for the Student Performance dataset, which contains Mathematics and Portuguese course records [18]. However, models that strongly depend on prior grades may have limited usefulness for early intervention when such grades are unavailable or when prediction is required across different academic domains.

Existing studies can be grouped into four broad categories. The first group focuses on centralized EDM, where student records are pooled and models are trained under a single-domain assumption. The second group emphasizes interpretable prediction using regression, decision trees, or post-hoc explanation tools. The third group investigates high-performing ensemble or neural models, but often assumes centralized access to student data. The fourth group explores privacy-preserving or federated learning, but with limited attention to subject-level domain shift and client-level personalization. This study is positioned at the intersection of these categories by comparing centralized learning, cross-domain generalization, and personalized federated learning under a unified evaluation protocol.

Recent literature emphasizes the importance of explainable and fair AI in academic prediction systems. Elastic Net Regression [14] is useful because it combines L1 and L2 regularization, supporting variable selection while handling multicollinearity. Support Vector Regression (SVR) [15] captures non-linear relationships through kernel functions, while tree-based ensemble algorithms such as XGBoost [21] are effective for structured tabular prediction tasks. In student-performance prediction, ensemble models have frequently shown strong predictive performance, especially when the data contain non-linear feature interactions [17], [4], [22]. However, these models are less transparent than linear models and often

require post-hoc interpretation methods such as SHAP [12] or LIME [13]. Since the present study does not include a separate SHAP experiment in the results section, explainability is treated here as a contextual motivation rather than a claimed experimental contribution.

Beyond centralized modeling, cross-domain generalization remains an important but less frequently tested problem in EDM. A model trained on one academic subject may not generalize well to another subject because the source and target domains can differ in feature distributions, target distributions, and predictor–outcome relationships. In this study, the Portuguese $\rightarrow$ Mathematics experiment is therefore treated as a cross-domain transfer setting under possible covariate shift and representation mismatch. This framing is important because poor cross-domain performance does not necessarily imply that a model is weak; rather, it may indicate that relationships learned from the source domain are not stable in the target domain.

Federated learning (FL) has emerged as a privacy-preserving learning paradigm in which multiple clients collaboratively train a model without centralizing raw data. FedProx was proposed to address statistical and systems heterogeneity in federated networks by adding a proximal term to stabilize local optimization [19]. FedBN was later introduced to handle non-IID feature distributions by keeping batch-normalization statistics local while sharing other model parameters [20]. These two methods are relevant to educational analytics because different institutions, departments, or subject groups may have heterogeneous student populations and restricted data-sharing policies. However, most existing educational prediction studies still focus on centralized settings, while fewer studies jointly evaluate centralized learning, cross-domain transfer, and personalized federated learning.

The literature, therefore, reveals three main gaps. First, many student-performance prediction studies evaluate models within a centralized single-domain setting, which may not reflect real deployment conditions across different subjects or institutions. Second, cross-domain validation is still limited, even though academic models may be applied to cohorts or subjects that differ from the training domain. Third, federated learning has not been sufficiently examined for personalized student-performance prediction under subject-level heterogeneity. This study addresses these gaps by evaluating centralized learning, Portuguese $\rightarrow$ Mathematics cross-domain transfer, and a personalized federated learning framework using consistent regression metrics.

A. Student Outcome Prediction Using Regression, Decision Trees, and Neural Networks

Regression-based models have long been used in EDM to quantify the effects of predictors such as study time, attendance, parental education, and previous grades [23], [16]. These models remain valuable because their coefficients provide relatively transparent evidence about predictor importance. However, linear models may struggle when relationships between variables are non-linear or highly interactive [24]. Decision-tree methods address some of these limitations by producing human-readable rules and modeling interactions between categorical and continuous variables [23], [25]. Neural networks, including multilayer perceptrons, can learn higher-order interactions when sufficient data are available, but their internal decision mechanisms are typically less transparent to educational stakeholders [26].

B. Ensemble Tree Methods and XGBoost

Ensemble tree algorithms, such as random forest and gradient boosting, are widely used for tabular prediction because they combine multiple weak learners to improve accuracy and robustness. XGBoost is a scalable gradient-boosting framework that has been widely adopted for structured prediction tasks [21]. In educational datasets, ensemble methods can capture non-linear relationships and feature interactions more effectively than simple linear models [22]. Nevertheless, their predictions may require post-hoc explanation methods when used in decision-support settings, particularly where teachers or administrators need to understand why a student is predicted to be at risk.

C. Interpretability and Ethical Considerations in Educational AI

Interpretability is important in educational applications because predictive models may influence interventions, resource allocation, or academic support decisions. Opaque models can raise concerns related to fairness, accountability, and trust [27]. Rudin [28] argues that high-stakes applications should prefer interpretable models where possible rather than relying only on post-hoc explanations of black-box models. In education, this issue is especially relevant because teachers and institutional decision-makers may be reluctant to use AI-based recommendations if the reasoning behind predictions is unclear [29]. Therefore, model performance should be interpreted alongside transparency, practical usability, and deployment constraints.

D. Feature Engineering and Representation Learning

Educational data often include heterogeneous demographic, behavioral, social, and academic variables. Feature engineering uses domain knowledge to transform these variables into meaningful predictors, such as interaction terms or normalized behavioral indicators. Regularized regression models such as Elastic Net can support this process by selecting relevant variables while controlling multicollinearity [14], [30]. This makes them useful as transparent baselines in educational analytics.

In contrast, representation-learning approaches such as ensemble models and neural networks learn complex patterns directly from data. Models such as XGBoost and MLPs can improve predictive performance by modeling non-linear interactions, but this often reduces direct interpretability [21], [22]. Explainable AI methods such as SHAP and LIME can partially bridge this gap by estimating feature contributions for complex models [12], [13]. However, because the present study focuses primarily on comparing learning paradigms rather than conducting a full XAI analysis, interpretability is discussed as a supporting consideration rather than as a separate experimental claim.

E. Federated Learning and Personalization

Federated learning is suitable for settings where data are distributed across institutions and cannot be centrally pooled. Standard FL can still suffer under non-IID data because client distributions may differ substantially. FedProx addresses this problem by regularizing local updates so that client models do not drift too far from the global model during local training [19]. FedBN addresses feature-shift non-IID data by preserving local batch-normalization statistics, which allows clients to maintain domain-specific normalization behavior [20]. In student-performance prediction, these mechanisms are relevant because different subjects, schools, or departments may produce different data distributions. Personalization further allows each client to adapt the global model to its local data distribution, which is important when a single global model cannot fully represent all clients.

F. Model Comparison, Trade-Offs, and Research Gap

Prior studies show a recurring trade-off between interpretable models, which are easier to justify but may have lower predictive capacity, and complex models, which may provide stronger prediction but lower transparency. However, fewer studies evaluate this trade-off together with cross-domain generalization and federated personalization. This gap is important because real educational analytics systems may face three simultaneous constraints: limited data sharing, distributional differences across domains, and the need for reliable local predictions.

This study, therefore, compares centralized, cross-domain, and federated learning settings using the same student-performance dataset family and the same regression metrics. The centralized experiment evaluates predictive performance when data pooling is possible. The cross-domain experiment tests whether a model trained on Portuguese records can generalize to Mathematics records. The federated experiment evaluates whether collaborative training with personalization can improve local adaptation while preserving data locality.

Table I summarizes the positioning of this study relative to prior work. Existing studies commonly focus on centralized accuracy, interpretability, cross-domain transfer, or privacy-preserving learning in isolation. In contrast, this work evaluates these issues together by comparing centralized learning, Portuguese $\rightarrow$ Mathematics cross-domain generalization, and personalized federated learning under a consistent experimental protocol. This structure directly addresses the need for educational prediction models that are accurate, privacy-aware, and robust to domain heterogeneity.

III. Methodology

The methodological framework of this study, illustrated in Fig. 1, was designed to systematically evaluate and compare machine learning approaches for predicting student academic performance. The framework is organized into three major experimental phases: **1)** centralized regression and ensemble learning baselines, **2)** cross-domain generalization across subject-specific datasets, and **3)** a privacy-preserving federated learning framework integrating *FedProx* and *FedBN*. All experiments were conducted using the UCI Student Performance dataset [18], which comprises demographic, behavioral, and academic attributes of secondary school students enrolled in Portuguese and Mathematics courses. Within this framework, a comparative modeling strategy was employed to predict students' final academic outcomes in Mathematics (`G3_mat`) and Portuguese (`G3_por`) using three representative machine learning paradigms: Elastic Net Regression, Multi-Layer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost) were selected to balance predictive complexity, interpretability, and robustness across heterogeneous educational datasets. Elastic Net offers transparent linear modeling with resilience to multicollinearity, making it suitable for interpretable academic performance analysis [14], [30]. MLP is employed to capture complex non-linear relationships among educational indicators [31], [32], while XGBoost delivers strong predictive capability through ensemble-based learning and feature interaction modeling [21]. Beyond centralized evaluation, these models are further integrated within a Federated learning framework, enabling privacy-preserving collaborative training across distributed institutional datasets without direct data sharing. This federated setup facilitates the assessment of global generalization, domain heterogeneity, and personalization trade-offs, thereby extending traditional educational data mining toward scalable, secure, and real-world multi-institutional analytics [23], [33].

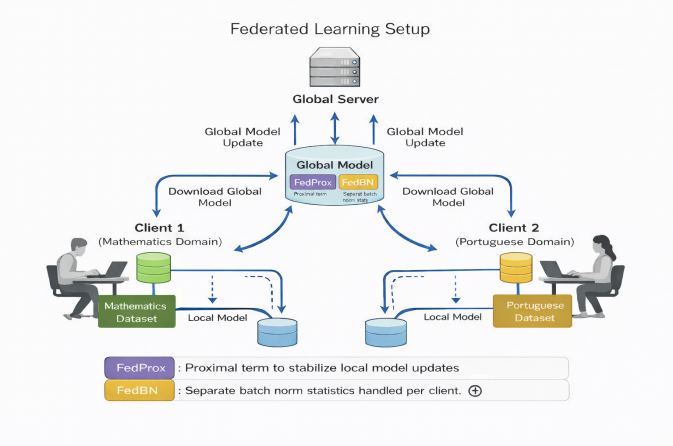

Fig. 1. Overall methodological framework for centralized learning, cross-domain generalization, and personalized federated learning.

A. Dataset Description

1) Data preprocessing and experimental setup: A unified, reproducible preprocessing framework was employed across three experimental settings to ensure fair comparisons across centralized, cross-domain, and federated learning paradigms.

a) Experiment 1 (Combined dataset): The Portuguese and Mathematics student datasets were merged using their shared feature space, excluding intermediate and final grade variables ($G1$, $G2$, and $G3$). A unified target variable, $G3_{\text{combined}}$, was constructed by averaging the final grades from both domains, and instances with missing target values were removed. The resulting dataset was separated into input features and target labels. A preprocessing pipeline based on a `ColumnTransformer` was applied, with numerical features imputed using mean substitution and categorical features transformed via one-hot encoding with unseen-category

TABLE I. COMPARISON OF PRIOR STUDIES ON STUDENT PERFORMANCE PREDICTION AND THE POSITIONING OF THIS WORK

Study Category	Learning Paradigm	Model Types	Key Limitations in Prior Work	Positioning of this Study
Early EDM Studies [16]	Centralized	Decision Trees, Random Forests, Neural Networks, Regression	Mostly single-domain evaluation; strong dependence on prior grades	Uses the same dataset family but evaluates centralized, cross-domain, and federated settings
Regression-Based EDM [14], [23]	Centralized	Linear Regression, Elastic Net	High interpretability but limited capacity for complex non-linear relationships	Uses interpretable regression models as baselines against non-linear models
Ensemble and Neural Models [21], [22]	Centralized	XGBoost, Random Forests, MLPs	Strong prediction but reduced transparency and limited deployment analysis	Compares high-capacity centralized models with cross-domain and federated alternatives
Explainable AI in Education [12], [13]	Post-hoc XAI	SHAP, LIME	Explanations often studied separately from privacy and domain-shift constraints	Uses explainability literature to motivate transparency concerns, without claiming a separate XAI experiment
Cross-Domain Learning in EDM	Centralized Transfer	Various ML Models	Generalization across subjects is often assumed rather than explicitly tested	Evaluates Portuguese → Mathematics transfer and identifies domain-shift effects
Federated Learning [19], [20]	Federated / Personalized	FedProx, FedBN, Personalized FL	Limited use in student-performance prediction; personalization often underexplored	Evaluates personalized FL for heterogeneous educational data while keeping raw data local
This Work	**Centralized, Cross-Domain, Federated**	**Elastic Net, SVR, XGBoost, MLP, FL Neural Models**	**Need for unified comparison under privacy and domain-shift constraints**	**Provides a unified evaluation showing strong centralized baselines, weak direct cross-domain transfer, and improved local adaptation through personalized FL**

handling. The processed data were subsequently split into training and test sets at a 70:30 ratio and integrated into end-to-end pipelines for Linear Regression, Support Vector Regression (SVR), Elastic Net, Multi-Layer Perceptron (MLP), and XGBoost models.

b) Experiment 2 (Cross-domain evaluation): To simulate a realistic source–target transfer learning scenario, the Portuguese and Mathematics datasets were treated as independent domains. Feature matrices and corresponding targets were extracted independently after excluding grade-related variables. In this setting, the preprocessing transformer was trained exclusively on the Portuguese dataset and applied to the Mathematics dataset without refitting. This design ensured that numerical imputation statistics and categorical encoding schemes were learned solely from the source domain, enabling an unbiased assessment of model generalization under domain shift. The resulting preprocessing pipeline was consistently used with XGBoost, SVR, Elastic Net, and MLP models.

c) Experiment 3 (Federated learning simulation): For federated learning, a global preprocessing strategy was adopted to guarantee feature consistency across decentralized clients while preserving data locality. Feature matrices from both domains were concatenated only for fitting a shared preprocessing transformer, without exposing labels or performing centralized model training. The federated preprocessor applied mean imputation followed by standardization for numerical attributes, which is essential for stable neural network optimization, and one-hot encoding for categorical features. The fitted transformer was then applied independently to each client dataset, producing locally encoded feature representations with identical dimensionality. These encoded inputs defined the neural network input space and were subsequently used within a federated optimization framework that employed FedProx with Batch Normalization personalization (FedBN), enabling privacy-preserving collaborative learning across heterogeneous educational data distributions.

B. Evaluation Metrics

All experiments were evaluated using the Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to ensure consistent and fair comparison across centralized, cross-domain, and federated learning setups. The R^2 metric measures the proportion of variance in the target variable explained by the model and is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{1}$$

where, y_i denotes the true target value, $\hat{y}_i$ is the corresponding model prediction, and $\bar{y}$ represents the mean of the observed targets. The Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors and is given by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{2}$$

while the Root Mean Squared Error (RMSE) emphasizes larger errors through quadratic penalization and is computed as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{3}$$

In regression-based evaluation, no single performance metric is universally optimal, as each metric captures different aspects of model behavior. Widely used measures such as the Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) provide complementary perspectives on predictive accuracy and error distribution. Recent studies highlight important nuances in metric selection. Chicco et al. [34] argue that R^2 is often more

informative than error-based metrics, as it offers a holistic assessment of model fit by quantifying the proportion of variance explained, whereas MAE and RMSE may exhibit limitations in interpretability when used independently. Conversely, Plevris et al. [35] emphasize the practical strengths of error-based measures, noting that MAE is robust to noisy datasets and less sensitive to outliers, while RMSE provides a balance between robustness and precision by assigning higher penalties to larger errors. Additionally, mean squared error (MSE) is particularly effective when minimizing the impact of substantial deviations is a primary objective. More recently, Dumre et al. [36] stress that evaluation metric selection should be aligned with specific research objectives and dataset characteristics, as different applications prioritize overall model fit, outlier sensitivity, or predictive precision. Accordingly, employing a multi-metric evaluation strategy is recommended to ensure a comprehensive and reliable assessment of regression model performance.

For federated learning experiments, performance was evaluated at both the global and local levels. *Global model performance* measures the predictive capability of the aggregated federated model and reflects the effectiveness of collaborative knowledge sharing across clients. In contrast, *personalized local performance* evaluates each client's locally adapted model on its own data distribution, capturing domain-specific learning and personalization effects. Reporting both metrics enables a comprehensive analysis of the trade-off between global generalization and client-level adaptation in heterogeneous federated environments.

IV. Results and Discussion

A. Overview

This section presents and analyzes the results from three experimental phases:

- Centralized regression and ensemble baselines,
- Cross-domain generalization, and
- Federated learning (FedProx + FedBN).

Performance was evaluated using the Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The objective was to assess not only the predictive accuracy of the models but also their interpretability, robustness, and scalability across data distributions.

This study investigates multiple machine learning paradigms for predicting student academic performance, including centralized learning, cross-domain generalization, and federated learning with personalization. The primary objective across all experiments is to predict final student grades ($G3$) using Portuguese and Mathematics student datasets under varying data availability and distributional assumptions.

B. Experiment 1: Centralized Learning on Combined Dataset

In the first experiment, all student records from the Portuguese and Mathematics datasets were centrally aggregated, and several regression models were evaluated, including Linear Regression, XGBoost, Support Vector Regression (SVR), Elastic Net, and a Multi-Layer Perceptron (MLP) Regressor (see Table II).

Among all evaluated models, XGBoost achieved the strongest predictive performance, attaining an R^2 score of 0.8308, along with a Mean Absolute Error (MAE) of 0.9356 and a Root Mean Squared Error (RMSE) of 1.5558. This superior performance highlights the effectiveness of tree-based ensemble methods in capturing complex, non-linear relationships present in student performance data when a rich, centrally available dataset is accessible.

The MLP Regressor also demonstrated competitive performance, achieving an R^2 of 0.7830, MAE of 1.1415, and RMSE of 1.7618. This result further suggests the presence of non-linear dependencies among educational features that neural networks are well suited to model.

In contrast, simpler models such as Linear Regression ($R^2 = 0.7681$), Elastic Net ($R^2 = 0.7441$), and SVR ($R^2 = 0.7229$) produced respectable, but comparatively lower predictive accuracy. While these models maintain interpretability and stability, their limited representational capacity restricts their ability to fully model the underlying complexity of student achievement patterns.

1) Conclusion for Experiment 1: When data can be centrally aggregated, models capable of learning non-linear feature interactions—particularly XGBoost and MLP—provide robust and highly accurate predictions, making them well suited for centralized student performance prediction tasks.

TABLE II. Performance of Centralized Models on the Combined Portuguese and Mathematics Student Dataset.

Model	Training / Evaluation	R^2	MAE	RMSE
Linear Regression	Combined (POR + MAT)	0.7681	1.1535	1.8212
XGBoost	Combined (POR + MAT)	**0.8308**	**0.9356**	**1.5558**
SVR	Combined (POR + MAT)	0.7229	1.1451	1.9908
Elastic Net	Combined (POR + MAT)	0.7441	1.1813	1.9131
MLP Regressor	Combined (POR + MAT)	0.7830	1.1415	1.7618

C. Experiment 2: Cross-Domain Generalization

The second experiment examined the feasibility of cross-domain generalization by training models exclusively on the Portuguese dataset and evaluating them on the Mathematics dataset. This scenario simulates real-world conditions where models trained in one educational context are applied to a different institutional or curricular setting (see Table III).

Across all evaluated models, performance deteriorated substantially. XGBoost, despite its strong centralized performance, achieved only a modest R^2 score of 0.1443, with MAE and RMSE values of 3.0015 and 4.2327, respectively. More notably, SVR ($R^2 = -0.0445$), Elastic Net ($R^2 = -0.0602$), and the MLP Regressor ($R^2 = -0.0701$) all yielded negative R^2 scores, indicating predictive performance worse than a naive mean-based baseline.

1) Conclusion for Experiment 2: The severe and consistent degradation in performance across all models confirms the presence of a significant domain shift between the Portuguese and Mathematics student datasets. These findings demonstrate that direct cross-domain deployment of machine learning models without adaptation is highly ineffective and leads to unreliable predictions.

TABLE III. CROSS-DOMAIN PERFORMANCE OF MODELS TRAINED ON THE PORTUGUESE DATA AND TESTED ON MATHEMATICS DATA.

Model	R^2	MAE	RMSE
XGBoost	0.1443	3.0015	4.2327
SVR	-0.0445	3.3561	4.6764
Elastic Net	-0.0602	3.4408	4.7113
MLP Regressor	-0.0701	3.2265	4.7333

D. Experiment 3: Federated Learning with Personalization

The third experiment explored a federated learning (FL) framework combining FedProx regularization with FedBN-style local batch-normalization personalization, followed by local model personalization (see Table IV).

The globally aggregated federated model achieved a moderate R^2 score of 0.6694, with MAE of 1.5508 and RMSE of 2.2211 on the combined evaluation set. While this performance surpasses direct cross-domain application, it remains lower than the best centralized model, indicating that a single global federated model may not fully represent heterogeneous local data distributions.

In contrast, the personalized federated models exhibited substantial performance gains. The personalized Portuguese client achieved an R^2 score of 0.8564, MAE of 0.9286, and RMSE of 1.2233, while the personalized Mathematics client achieved an R^2 score of 0.7805, MAE of 1.6149, and RMSE of 2.1437. These results indicate that local personalization substantially improves client-specific predictive performance compared with the global federated model. However, direct comparison with centralized learning should be interpreted cautiously because the evaluation scopes differ [see Fig. (2) to Fig. (4)].

1) Conclusion for Experiment 3: Federated learning combined with local personalization proves highly effective for distributed and heterogeneous educational data. The global model provides a strong initialization, while local fine-tuning enables clients to adapt to their unique data distributions, resulting in improved client-specific performance compared with the global federated model.

TABLE IV. PERFORMANCE OF FEDERATED LEARNING MODELS USING FEDPROX AND FEDBN WITH GLOBAL AND PERSONALIZED EVALUATION.

Model Variant	Evaluation Scope	R^2	MAE	RMSE
Global FL Model (FedProx + FedBN)	Combined (Global)	0.6694	1.5508	2.2211
Personalized FL Model	Portuguese Client	**0.8564**	**0.9286**	**1.2233**
Personalized FL Model	Mathematics Client	0.7805	1.6149	2.1437

V. DISCUSSION AND CONCLUSION

This study provides several important insights into student performance prediction under centralized, cross-domain, and federated learning paradigms. First, the results demonstrate that when centralized data pooling is feasible, models with higher representational capacity—particularly ensemble and neural approaches—achieve superior predictive accuracy compared to linear and regularized regression models. This highlights the importance of modeling non-linear interactions among socio-academic features in educational datasets.

Second, the consistently poor performance observed across domains confirms that domain shift between academic subjects is a fundamental challenge. Models trained on Portuguese data fail to generalize effectively to Mathematics outcomes, even when trained with increasing amounts of source-domain data. These findings indicate that naïve cross-domain deployment without adaptation leads to unreliable predictions and is unsuitable for real-world educational settings. Most importantly, the federated learning experiments demonstrate that personalization is critical for learning in heterogeneous, privacy-constrained environments. While the global federated model captures shared patterns across institutions, client-level personalization enables effective adaptation to local data distributions. Personalized federated models outperform the global federated model and achieve competitive local performance, illustrating the value of combining shared global learning with client-specific adaptation. Overall, these findings suggest that federated learning with personalization offers a promising privacy-preserving alternative to centralized modeling, particularly when student data are distributed across heterogeneous academic domains.

A. Learning Curve Analysis

Learning curve analysis was used to further examine convergence behavior and generalization trends across the evaluated learning paradigms. In centralized learning, model performance stabilizes as training data increases, with non-linear models benefiting most from additional data, indicating effective convergence under single-domain conditions.

In contrast, cross-domain learning exhibits persistent generalization gaps, with test performance remaining low or even negative despite increased source-domain data. This behavior confirms that domain mismatch, rather than data scarcity, is the primary cause of performance degradation in cross-domain settings.

Federated learning shows consistent improvement across communication rounds, demonstrating effective collaborative optimization. Personalized models converge faster and achieve higher performance than the global federated model, reinforcing the importance of heterogeneity-aware learning strategies in decentralized environments.

B. Key Implications

The experimental results lead to three central implications:

- Centralized learning remains effective when data sharing is possible but is limited by privacy constraints and domain variability.
- Direct cross-domain transfer across academic subjects is ineffective without explicit adaptation mechanisms.
- Federated learning with client-level personalization provides a practical and high-performing solution for multi-institutional educational analytics.

Overall, the findings show that centralized learning is effective when data pooling is possible, but its practical use may be limited by institutional privacy constraints. Direct cross-domain transfer from Portuguese to Mathematics produced weak generalization, indicating that subject-level domain shift

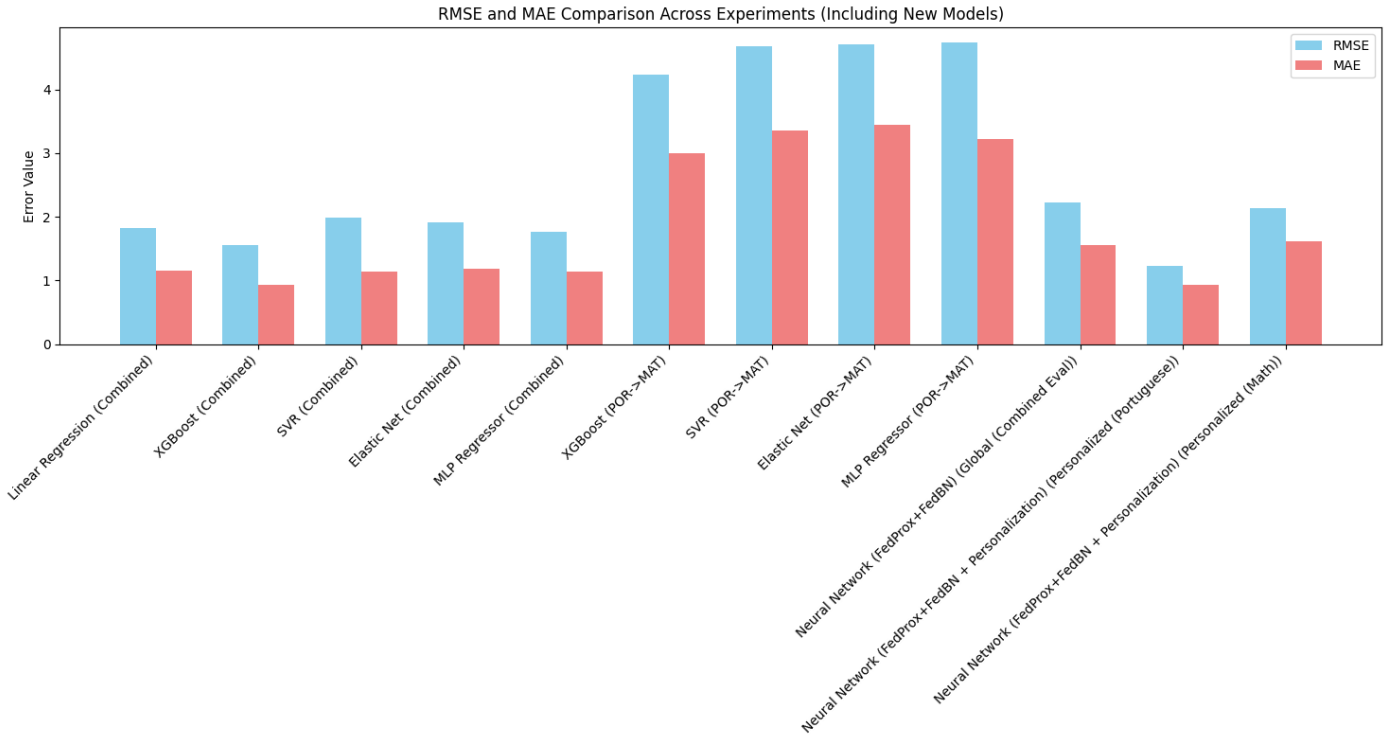

Fig. 2. RMSE and MAE comparison across centralized, cross-domain, and federated learning models.

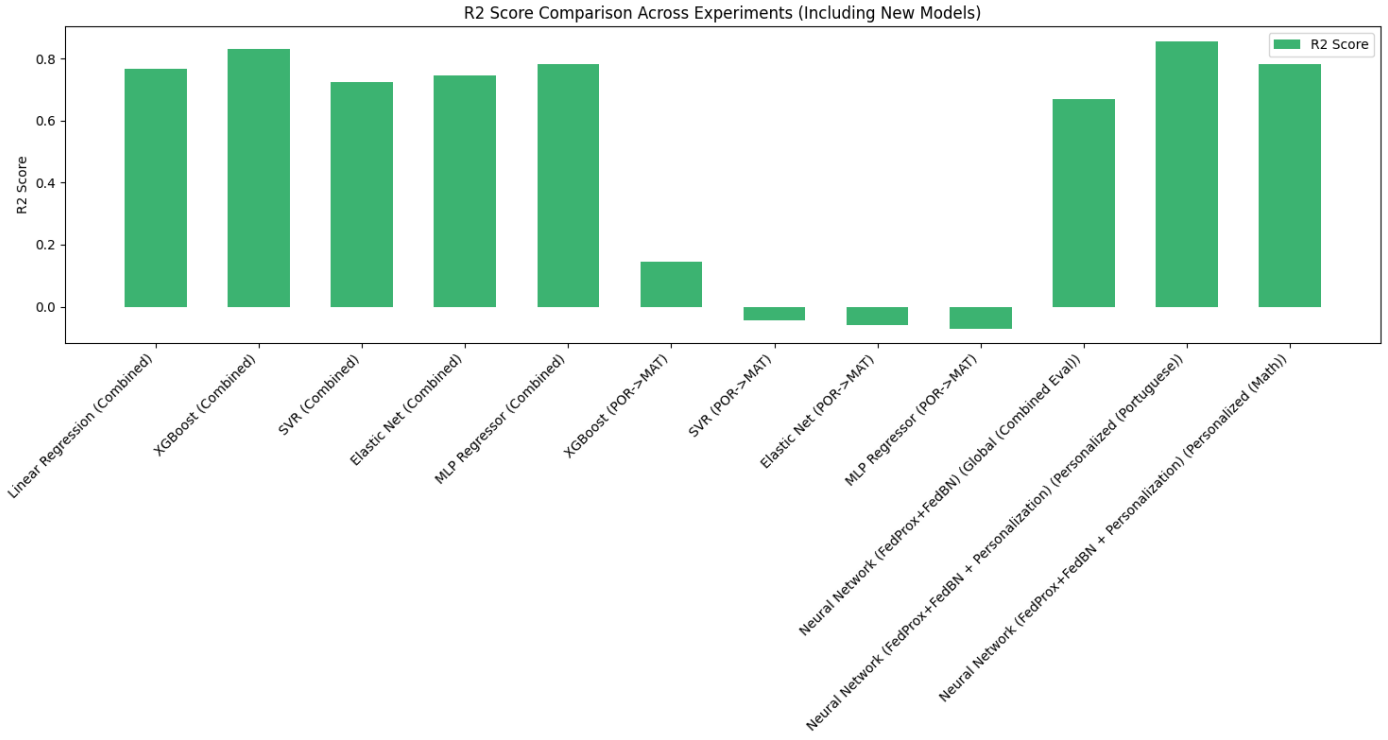

Fig. 3. R^2score comparison across centralized, cross-domain, and personalized federated learning models.

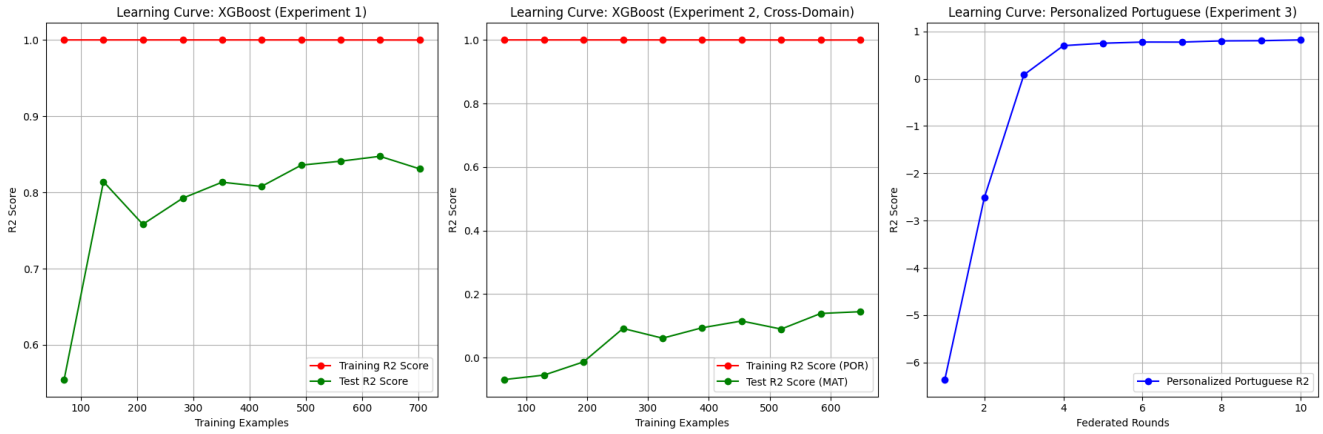

Fig. 4. Learning-curve comparison for centralized XGBoost, cross-domain transfer, and personalized federated learning.

must be explicitly addressed. The federated learning results show that a global model can capture shared patterns across domains, while client-level personalization improves local predictive performance. However, these findings are limited to the UCI Student Performance dataset and a simulated two-client federated setting. Future work should evaluate larger multi-institutional datasets, repeat experiments across multiple random seeds, report statistical confidence intervals, and measure communication cost under realistic federated deployment conditions.

REFERENCES

[1] M. N. Yakubu and A. M. Abubakar, "Applying machine learning approach to predict students' performance in higher educational institutions," *Kybernetes*, vol. 51, no. 2, pp. 916–934, 2022.

[2] W. Ahmed, M. A. Wani, P. Plawiak, S. Meshoul, A. Mahmoud, and M. Hammad, "Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions," *Scientific Reports*, vol. 15, no. 1, p. 26879, 2025.

[3] Y. Zhang, Y. Zhang, and J. Zhang, "Educational data mining techniques for student performance prediction: A review," *Frontiers in Psychology*, vol. 12, 2021.

[4] I. Issah, M. Asante, and S. Amankwah, "A systematic review of machine learning techniques for student performance prediction," *Heliyon*, vol. 9, no. 5, 2023.

[5] X. Bai, F. Zhang, J. Li, T. Guo, A. Aziz, A. Jin, and F. Xia, "Educational big data: Predictions, applications and challenges," *Big Data Research*, vol. 26, p. 100270, 2021.

[6] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student'performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, p. 552, 2021.

[7] O. P. Ojajuni, F. Ayeni, O. Akodu, F. Ekanoye, S. Adewole, T. Ayo, S. Misra, and V. W. A. Mbarika, "Predicting student academic performance using machine learning," in *Communication Systems and Applications*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237550930

[8] A. Razaque and A. Alajlan, "Supervised machine learning model-based approach for performance prediction of students," *Journal of Computer Science*, vol. 16, no. 8, pp. 1150–1162, 2020.

[9] M. Arifin, W. Widowati, F. Farikhin, and G. Gudnanto, "A regression model and a combination of academic and non-academic features to predict student academic performance," *TEM Journal*, vol. 12, no. 2, p. 855, 2023.

[10] A. Mahmood, Z. H. Azizul, M. Zakariah, S. B. Belhaouari, A. Altameem, R. Ramli, A. S. Almazyad, M. L. M. Kiah, and S. R. Azzuhri, "Implementing federated learning over vpn-based wireless backhaul networks for healthcare systems," *PeerJ Computer Science*, vol. 10, p. e2422, 2024.

[11] D. Subramanian, A. Ajitha, and S. S. Maidin, "Unveiling hybrid model with naive bayes, deep learning, logistic regression for predicting customer churn and boost retention," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 1379–1391, 2025.

[12] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017. [Online]. Available: http://arxiv.org/abs/1705.07874

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[15] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[16] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," https://www.researchgate.net/publication/228780408, 2008.

[17] M. Rashid, F. A. Khan, A. Alqahtani, and M. Hameed, "Classifying and predicting students' performance using popular algorithms," *Journal of Computer Science*, vol. 15, no. 8, pp. 1179–1190, 2019.

[18] UCI Machine Learning Repository, "Student performance dataset," 2014, introductory paper: Cortez and Silva, 2008.

[19] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, vol. 2, 2020, pp. 429–450.

[20] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," in *International Conference on Learning Representations*, 2021.

[21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[22] H. Sahlaoui, E. A. A. Alaoui, A. Nayyar, S. Agoujil, and M. M. Jaber, "Predicting and interpreting student performance using ensemble models and shapley additive explanations," *IEEE Access*, vol. 9, pp. 152 688–152 703, 2021.

[23] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 2010.

[24] M. A. Babyak, "What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models," *Psychosomatic Medicine*, 2004.

[25] M. Hoq, P. Brusilovsky, and B. Akram, "Analysis of an explainable student performance prediction model in an introductory programming course," 2023.

[26] S. G. Mundhe, S. Y. Gaikwad, and A. Professor, "Performance prediction in educational data mining using neural network," *International Journal of Innovative Research in Advanced Engineering*, vol. 8, pp. 582–585, 2019.

[27] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive models for educational purposes: A systematic review," *Big Data and Cognitive Computing*, 2024.

[28] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, 2019.

[29] M. Lucas, Y. Zhang, P. Bem-haja, and P. N. Vicente, "The interplay between teachers' trust in artificial intelligence and digital competence," *Education and Information Technologies*, vol. 29, no. 17, pp. 22 991–23 010, 2024.

[30] J. K. Tay, N. Aghaeepour, T. Hastie, and R. Tibshirani, "Feature-weighted elastic net: Using features of features for better prediction," *Statistica Sinica*, vol. 33, no. 1, pp. 259–279, 2023.

[31] Y. Liu, C. Zhao, and Y. Huang, "A combined model for multivariate time series forecasting based on mlp-feedforward attention-lstm," *IEEE Access*, vol. 10, pp. 88 644–88 654, 2022.

[32] C. Enăchescu, "Approximation capabilities of neural networks," *Journal of Numerical Analysis, Industrial and Applied Mathematics*, vol. 3, no. 4, pp. 221–230, 2008.

[33] E. Kalita *et al.*, "Educational data mining: a 10-year review," *Information Retrieval Journal*, 2025.

[34] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *Peerj computer science*, vol. 7, p. e623, 2021.

[35] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," 2022.

[36] P. Dumre, S. Bhattarai, and H. K. Shashikala, "Optimizing linear regression models: A comparative study of error metrics," *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp. 1856–1861, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:275789357