

AFFORDLEARN: EXPLAINABLE STACKED ENSEMBLE REGRESSION FOR INTERNATIONAL EDUCATION COST PREDICTION

Zeeshan Raza¹, Huma Chaudhry², Muhammad Akmal³, Irfan Ul Haq⁴,
Muhammad Qadeer⁵, Misbah Akram⁶, Muhammad Usman Javeed^{*7}

¹Department of Computer Science, COMSATS University Islamabad (CUI), Sahiwal Campus, Sahiwal, Punjab,
Pakistan,

²School of Software Engineering, Minhaj University Lahore, Lahore, Pakistan,

³School of Software Engineering, Minhaj University Lahore, Lahore, Pakistan,

⁴Department of Computer and Information Sciences (DCIS), Pakistan Institute of Engineering and Applied Sciences
(PIEAS), Islamabad, Pakistan,

⁵Department of Computer Science, COMSATS University Islamabad (CUI), Sahiwal Campus, Sahiwal, Punjab,
Pakistan,

⁶School of Software Engineering, Minhaj University Lahore, Lahore, Pakistan,

⁷Department of Computer Science, COMSATS University Islamabad (CUI), Sahiwal Campus, Sahiwal, Punjab,
Pakistan,

¹zeeshan.raza@yahoo.com, ²huma.cs@mul.edu.pk, ³akmal.cs@mul.edu.pk, ⁴irfan@email.com

⁵kadeer0303iub@gmail.com, ⁶misbahakram.cs@mul.edu.pk, ⁷usmanjavveed@gmail.com

DOI: <https://doi.org/10.5281/zenodo.23032744>

Keywords

International Education Cost
Prediction, Machine Learning,
Ensemble Learning, Regression
Models, Forecasting

Article History

Received: 01 June 2026

Accepted: 23 August 2026

Published: 29 September 2026

Copyright @Author

Corresponding Author: *

Muhammad Usman Javeed

Abstract

This paper provides a comprehensive discussion of machine learning regression models in predicting the total cost of international education, based on a real-life dataset that includes country, program, level, tuition, accommodation, visa, insurance, and features of living expenses. A common set of preprocessing was used to achieve uniformity between the models and an extensive selection of algorithms was examined including Linear Regression, Ridge, Lasso, ElasticNet, Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost, and Multi-Layer Perceptron. Based on these standards, we introduce a new stacked ensemble model, AffordLearn, that combines XGBoost and Gradient Boosting as base learners and a Random Forest meta-learner. The suggested architecture represents a suitable combination of the boosting and bagging approaches to enhance the generalization and strength. As shown in experimental results ($R^2 = 0.9723$, $RMSE = 1.97$, $MAE = 1.46$), the predictive accuracy of AffordLearn was highest compared to all the individual and ensemble baselines. The results emphasize the usefulness of hybrid ensemble learning in improving the modeling of the non-linear and complex interaction between costs, thereby providing a consistent and understandable predictor of international education costs. The research is informative to the students, policymakers, and other institutions that may want to use data-driven instruments to plan education costs globally.

INTRODUCTION

The increasing nature of higher education has become an international issue as it has impacted millions of students in their dreams to study in a foreign country. The financial cost of undertaking international education comprises several mutually supportive aspects such as tuition expenses, living costs, visa and insurance prices as well as accommodation expenses. Collectively these are the total cost of education, which may vary broadly between nations, universities, and levels of programs. Such complexity usually makes students and families confused when making an educational investment, especially in areas whereby their local economic indicators and exchange rates are constantly changing. The cost prediction would be a crucial decision making aid in this dynamic environment among students, policy makers and educational institutions.

Traditional methods used for cost forecasting usually rely on the use of linear regression. Although these methods are interpretable and simplified, they are necessarily less capable of detecting nonlinear relationships between multiple variables. Linear models do not necessarily perform well in situations where they are used in the context of multivariate educational or economic data, where the predictor variables have complex interdependencies [1]. Variables such as inflation, exchange rates, institutional ratings, and geographic location may have nonlinear interactions that are not clearly captured by linear regression. Additionally, these models also assume a uniform data distribution and constant variance, which is rarely true for real education spending data.

To address such shortcomings, machine learning (ML) algorithms and data driven methods have become more adaptable. Machine learning algorithms like tree-based ensembles, boosting and neural regressors are able to learn latent data patterns automatically without having any rigid linearity assumptions. This enables efficient processing of categorical and numerical data commonly found in international education datasets, including both quantitative (such as

tuition and rent) and qualitative (such as country and educational attainment) indicators.

Ensemble models consistently outperform traditional regression methods on heterogeneous data, such as those used to explain college tuition and student spending [2]. Models like Random Forest and Gradient Boosting aggregate predictions from many weaker learners (e.g., decision trees) to produce robust and generalized models. Random Forests, based on the principle of bagging, reduce overfitting by averaging predictions across many randomly generated trees, while boosting methods iteratively refine predictions by minimizing residual errors. Both approaches can handle nonlinearity and high dimensional feature spaces effectively. Stacking multiple ensemble learners can further improve performance by integrating their complementary perspectives on the data [3].

Regardless of these developments, the use of ensemble and stacking models in education cost predicting is not fully developed. The earlier research has mostly concentrated on predicting the tuition fee or student performance or dropout prediction, as opposed to cost estimation. Neural networks have also shown potential, but typically require huge data sets and resources to stabilize generalization [4]. Furthermore, the explainability of deep neural networks is still an issue, especially when it comes to the financial sector where explanation is essential. Students need accurate predictions and a clear explanation of why some cost factors are more influential than others.

Recent developments in explainable artificial intelligence (XAI) have partially alleviated this problem by proposing tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). These models allow complex model predictions to be broken down into contributions by function, so stakeholders can understand the impact of tuition, rent, or program length on total costs. Ensemble learning combined with interpretability frameworks leads to improved accuracy and reliability of predictive models [5]. This hybrid method provides a workable middle ground

between model performance and transparency, important characteristics for making decisions in education cost planning.

Based on these findings, this study presents AffordLearn, a new stacked ensemble regression model specifically designed to predict international education costs. The proposed model combines XGBoost and Gradient Boosting as base learners, and a meta-learner is a random forest that performs the final aggregation. Each of the base learners captures a different part of the data: XGBoost is the model gradient-based learning and regularization; Gradient augmentation is the minimization of residual error based on sequential correction; and the random forest meta-learner is more robust and averages prediction models. This mixed format combines the concepts of boosting and bagging and results in an equal reduction in bias and variance.

In this research, comprehensive experiments were conducted on a real world dataset that includes country specific and program related attributes such as tuition fees, annual rent, living expenses, insurance, and visa costs. The models were evaluated using four standard metrics: the coefficient of determination R^2 , the root mean square error (RMSE), the mean absolute error (MAE), and the calculation time. Experimental results show that the proposed AffordLearn model achieved superior prediction accuracy with $R^2 = 0.9723$, $RMSE = 1.97$ and $MAE = 1.46$, outperforming established ensemble methods such as Random Forest, Extra Trees and XGBoost. These results confirm that integrating multiple ensemble learners through a meta-learning framework improves both prediction accuracy and generalization across heterogeneous datasets.

In addition to its performance benefits, the proposed model allows interpretability with post-hoc analysis via SHAP and LIME. These explainable tools show that features, including items related to tuition and accommodation, have the greatest influence on forecast results, while others, including visa and insurance fees, have secondary effects. Interpretability increases the value of the model for policy and financial

planning by providing practical information instead of black-box forecasts.

The main contributions of this article are as follows:

1. A comparative study of the conventional linear regressors, the tree-based learners and the advanced ensemble models in terms of the consistent evaluation criteria (R^2 , RMSE, MAE, and computation time).

2. A proposed stacked ensemble regression model, AffordLearn, built as an XGBoost and Gradient Boosting base learners and Random Forest as a meta-learner.

3. Use of interpretability frameworks (SHAP and LIME), to gain insights into the feature contributions.

The rest of this paper is organized as follows. Section 2 presents a comprehensive literature review on prior studies related to cost prediction, ensemble learning, and model interpretability. Section 3 outlines the research methodology, including preprocessing, feature engineering, and model training. Section 4 introduces the proposed stacked ensemble model, AffordLearn, explaining its architecture and parameter configuration. Section 5 presents the experimental results and performance comparisons of different regression models. Finally, Section 6 concludes the study and suggests directions for future work.

Literature Review

A literature review is conducted on previous research on cost prediction, machine learning-based regression, ensemble modeling, and model interpretability. It also helps in identifying the key contributions and research gaps that led to the creation of the proposed AffordLearn model.

To date, cost estimation and forecasting has been carried out in the healthcare, construction and project management sectors, and its use in the education sector is minimal. In cost prediction and estimation, various articles are published like Ma et al. [6] developed a machine learning model to predict stationary costs and achieved significantly better results than traditional regression models. Chen et al. [7] showed that

machine learning models are much more accurate than traditional approaches in predicting construction costs. In this sense, Zhang et al. [8] proposed a multi-task learning framework to predict medical costs and improve accuracy and interpretability. Mandolini et al. [9] developed a general machine learning cost modeling procedure, while Inan et al. [10] integrated gradient-based models to maximize the accuracy of project cost estimation. If we talk about education domain, Patel et al. [2] proposed an ML-based system to predict the total net price of higher education programs in the United States and found that ensemble approaches outperformed simple regression models.

Machine learning algorithms are now becoming the applications of choice in predictive cost modeling as they model non-linear relationships between complex data. Regression-based prediction is popular and is performed using traditional algorithms such as Gradient Boosting [11], Random Forest [12], and XGBoost [13]. Kablan et al [5] used stacked ensemble learning to make predictions about medical outcomes and found that these had better predictive accuracy than individual models. It was also observed by Lu et al. [3] and Sakhravi et al. [?] that stacking based ensembles were more reliable and robust in predicting hydrological and software costs. New developments in ensemble learning have focused on hybrid structures and Zhu et al. [14] presented a stacked ensemble optimizer that stacked several meta-learners to improve performance. Tahmouresi et al. [15] used ensemble learning to improve the spatial clustering of environmental data, and it was also found to have good generalization power. These results are consistent with the initial theoretical hypotheses of Friedman [11] and Breiman [12] that mixing weak learners increases accuracy and stability. Additionally, ensemble models are particularly preferred for cost forecasting tasks containing mixed numerical and categorical data. According to Bell et al. [16], ensemble diversity could be more effective than deeper models, while Mandolini et al. [9] discovered that managing the interaction of

features in an ensemble approach is essential for financial data. Taken together, these studies support the applicability of stacked ensemble learning for informed predictions of the cost of education.

There are also new developments in hybrid and stacked models to take cost forecasting to the next level. Inan et al. [10] used LSTM-based architectures to predict project cost, where the model showed better results in time series cost prediction compared to traditional regressors. When applied to construction, the ensemble framework developed by Habib et al. [17] combines Random Forest and Gradient Boosting predictors, enabling more predictable and accurate cost estimates on projects of different types. Qassem et al. [18] applied stacking sets to software cost estimation and their results showed that layered meta-learners can minimize both bias and variance. Naderalvojud et al. [19] studied improving prediction in noisy datasets using multiple baseline models, with more emphasis on ensemble diversity that contributes to robustness. Meanwhile, Chen et al. [20] conducted an empirical comparison in construction cost modeling and found that XGBoost and CatBoost models are far superior to linear models in R2 and RMSE.

Although ensemble models are highly predictive, their black box nature makes them less interpretable. Explainable artificial intelligence (XAI) techniques are becoming increasingly popular to solve this problem. SHAP (SHapley Additive Explanations) by Lundberg and Lee [21] is a model contributor that provides predictions from feature-level models. It is important to note that LIME (Local Interpretable Model-Agnostic Explanations) proposed by Ribeiro et al. [22] is intended to locally approximate complex models to make them easier to understand. XAI has been used to gain trust and transparency in education AI systems. As stated by Alamri and Alharbi [23], explainable models of student success were examined and transparency of the model was considered important. Linardatos et al. [24] presented an overview of interpretability methods,

while Lozano-Murcia et al. [25] used XAI techniques in financial forecasting to demonstrate the contribution of features to cost variation. Papenmeier et al. [26], Slack et al. [27], and Sokol and Flach [28] also examined the impact of model accuracy and explanation reliability on user confidence and decision confidence. Recent articles by Hammann and Wouters [29-30] have explored the trade-off between explainability and accuracy suggesting that interpretability can improve the ethical and responsible use of AI. According to Bell et al. [16], Transparency improves practical usability even in situations where accuracy is slightly reduced. The work of Zhu et al. [14] and Tahmouresi et al. [15] also showed that the combination of ensemble models and SHAP or LIME explanations ensures accuracy and interpretability, two crucial characteristics of reliable AI-based decision systems in education.

Recent studies have demonstrated the effectiveness of machine learning for cost prediction across healthcare, construction, manufacturing, and project management domains [31-66]. Ensemble and boosting methods have shown strong performance in modeling complex and nonlinear cost relationships, while stacking approaches can further improve predictive robustness [31,37]. These findings motivate the application of stacked ensemble regression to international education cost prediction, where multiple financial factors interact to determine overall expenditure.

Based on the literature review, several gaps in the research are identified:

- The majority of the previous research estimates the tuition or institutional cost without modeling the entire cost structure that encompass accommodation, visa, insurance, and costs of living.
- Although stacking has been applied successfully in other fields it is not well researched in the context of higher education cost prediction.
- Most of the machine learning studies focus on accuracy rather than transparency, something that restricts its use in policy-making and decision-making.

- The single-country or institutionally trained models are seldom tested in different education settings.

- Not many research works compare the traditional, ensemble and neural models on a systematic basis in terms of standard performance measures.

To address these gaps, this research integrates the latest ensemble regression models (XGBoost, Gradient Boosting, and Random Forest) into a meta-learning model, complemented by post-hoc interpretability techniques (SHAP and LIME). This method aims to provide high predictive capability and transparency in modeling the costs of international education.

Research Methodology

This paper has a systematic experiment procedure, as depicted in Figure 1, in the sense that it is aimed at testing multiple regression algorithms to predict the overall cost of education. The approach is made up of five consecutive steps which are dataset acquisition and visualization, feature engineering, model training, evaluation, and interpretation.

Dataset Acquisition and Visualization

The data used in this study was taken from the Kaggle platform [30]. The dataset was retained in its raw form with the sole attempt to preprocess it to conform to machine learning models.

The dataset is comprised of data about the cost of international education in various countries and programs. It comprises academic and non-academic cost items including tuition fee, accommodation, visa fee, insurance, and the estimated cost of living per year. The target variable, Total Cost, is the overall financial need to attend higher education and it is calculated as the amount of tuition and all costs involved in the program. The most important features are categorical and include Country, Program, and Level (Bachelor's, Master's, PhD), whereas the numerical features are tuition, rent, insurance, visa fees, and living costs. The data is well organized, and it has a minimal pre-processing.

Three exploratory data visualizations were developed in order to get a better insight into the distribution and the characteristics of the dataset before modeling.

Figure 2 offers a box plot of the total cost of education by country. The countries are ranked on the basis of median cost, which places great differences. The median costs in African countries like Guinea, Tunisia and Libya are the lowest (under \$50, 000) and the greatest in the United States with programs costing more than \$400,000. This points out the hegemonic nature of geographical location in influencing access to an education.

Figure 3 presents a scatter plot of annual tuition against annual living expenses (left axis and right axis respectively). The bubble size is used to indicate the duration of the program whereas the color relates to the living cost index. Most programs fall below the tuition amounts of less than \$20,000 and the living costs of less than 15,000, with exceptions existing like tuition that is much higher in the U.S. universities. This reveals

how tuition, cost of living, and duration of the program interact to get the total spending.

Figure 4 shows the median total cost of Bachelor, Master and PhD programs. Doctoral education is the most expensive because it takes the longest time and the final cost is higher than that of the Bachelor's and Master's degree course. This comparison depicts the effect of program level on the overall financial burden.

Feature Engineering

In order to train the model, the data was preprocessed to satisfy the requirements of the model training involving the use of suitable transformation procedures on both categorical and numeric variables. Categorical variables (*Country*, *Program*, and *Level*) were encoded using one-hot encoding to make sure that no artificial ordinal relationship is created. This increased the feature space and enabled machine learning algorithms to deal with categorical variables.

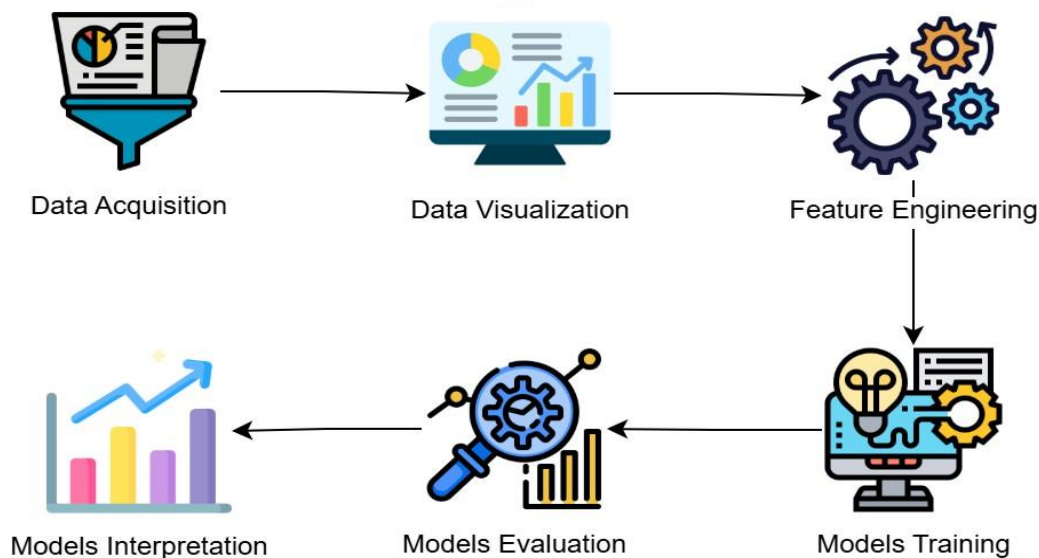


Figure 1. Research methodology

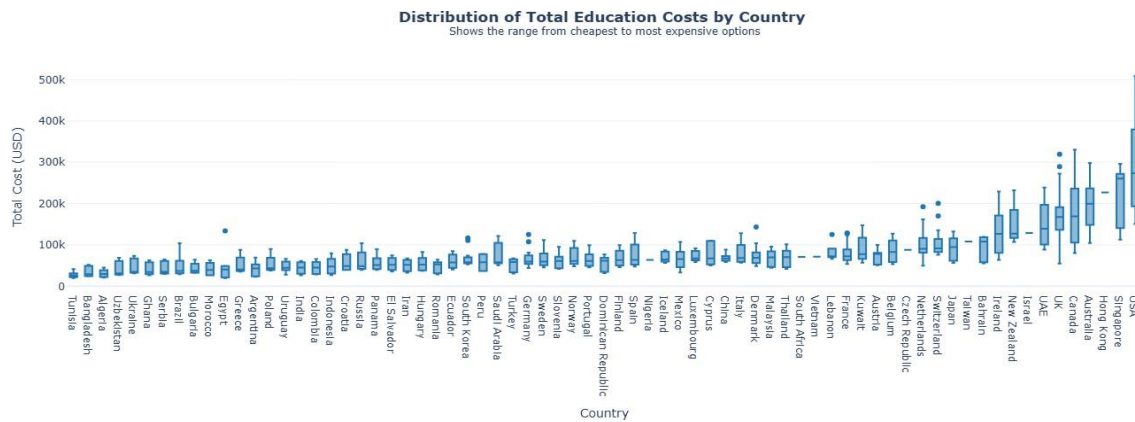


Figure 2. Distribution of Total Education Costs by Country

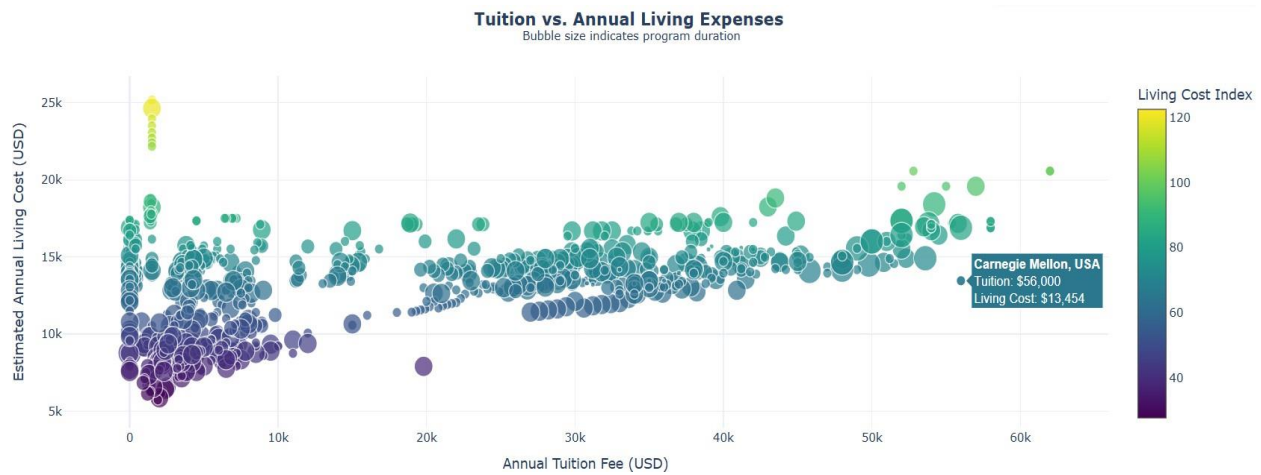


Figure 3. Tuition vs. Annual Living Expenses

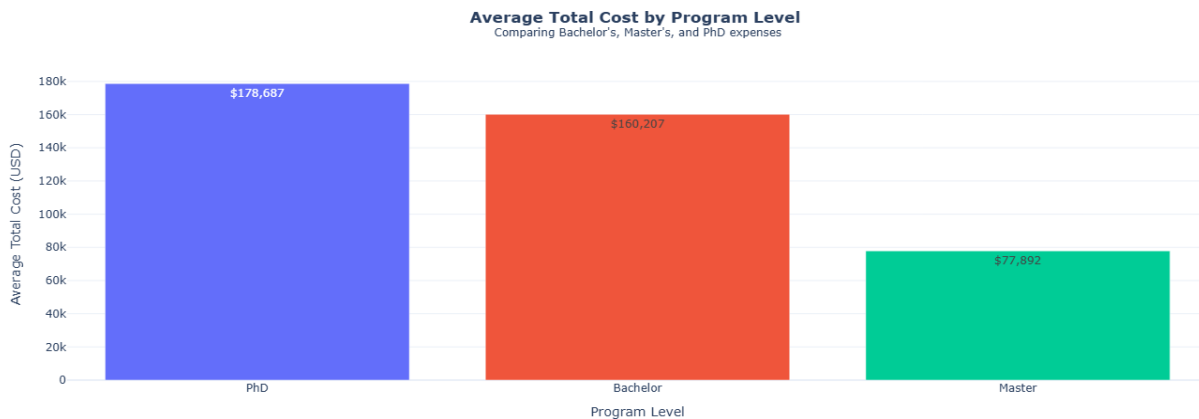


Figure 4. Average Total Cost by Program Level

Numerical variables such as tuition fees, annual rent, living expenses, insurance and visa cost were moreover standardized by using *StandardScaler* to remove the difference between scales and to enhance the convergence of gradient-based learners. Moreover, derived features reflecting the contextual variables, e.g. total cost of living index and duration-adjusted tuition, were added to boost predictive power. Missing values that existed were filled in with either mean imputation or mode imputation according to the type of data, so as not to lose any useful samples.

The preprocessing pipeline was made as a single operation in the sense that the same kind of transformations was done to both the training and the testing subsets. This is a replicable method, which ensured data integrity and reduced the chances of data leakage.

Models Training

Following the preprocessing, several regression algorithms were trained and tested in order to determine comparative performance metrics. These models comprised of the traditional linear models, which were Linear Regression, Ridge, Lasso, and ElasticNet, and non-linear models like Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost, Support Vector Regression (SVR), K-Nearest Neighbors (KNN) and AdaBoost. Table 1 explains the mathematical

formulation and key features of the regression models.

In order to have strong generalization, every model was trained after 80/20 train-test split. The randomized grid search was used to hyperparameter tuning, which optimized the following parameters: learning rate, number of estimators, maximum depth, and regularization coefficients.

The stacked ensemble strategy was used in the proposed model, which is named as *AffordLearn*. The first level had both XGBoost Regressor and Gradient Boosting Regressor as base learners and a second level had a Random Forest Regressor as the meta-learner. The meta-model was fed the predictions of the base learners and was taught to put them together in the best way possible thus it used the synergies of the two, boosting (reduced bias) and bagging (reduced variance).

Models Evaluation

In order to do a fair evaluation, the models were all trained on the same split of data and they were all tested under the same preprocessing conditions. The comparisons of performance were made in these metrics to determine the most effective trade-off between predictive power and computation power. The stacking ensemble design is beneficial as the proposed *AffordLearn* performed better in all evaluation criteria. The evaluation metrics of regression models are shown in Table 1.

Table 1. Evaluation Metrics of Regression Models

Model	Description	Mathematical Formulation	Key Features
Linear Regression (LR)	Fits a linear relationship between features and target.	$\hat{y} = X\beta + C$ $Lose: J(\beta) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$	Simple, interpretable, baseline regression method.
Decision Regressor (DT)	Recursive partitioning of feature space into regions with constant	$Prediction: \hat{y}_R = \frac{1}{ R } \sum_{i \in R} y_i$	Interpretable, prone to overfitting without pruning.

	prediction.		
Random Forest (RF)	Ensemble of Decision Trees trained on bootstrapped subsets. Prediction by averaging.	$\hat{y} = \frac{1}{M} \sum_{m=1}^M h_m(x)$	Reduces variance, robust, handles non-linear relations.
Extra Trees Regressor (ET)	Similar to RF but uses random splits for higher diversity.	$\hat{y} = \frac{1}{M} \sum_{m=1}^M h_m(x)$	Faster, reduces variance, more randomness.
Gradient Boosting Regressor (GBR)	Sequentially builds trees to correct errors of prior models.	$F_m(x) = F_{m-1}(x) + v h_m(x)$	Strong predictive performance, risk of overfitting if not tuned.
XGBoost Regressor	Optimized gradient regularization.	$L^t = \sum_i l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$	Highly efficient, regularized, widely used in competitions.
Support Vector Regressor (SVR)	Finds a regression hyperplane with margin of tolerance ϵ .	$\min_{w,b} \frac{1}{2} \ w\ ^2 \text{ s.t. } y_i - (w \cdot x_i + b) \leq \epsilon$	Effective in high-dimensional space, kernel trick.
Ridge Regression	Linear regression regularization.	$J(\beta) = \sum_i (y_i - \hat{y}_i)^2 + \lambda \ \beta\ _2^2$	Controls multicollinearity, reduces overfitting.
Lasso Regression	Linear regression regularization.	$J(\beta) = \sum_i (y_i - \hat{y}_i)^2 + \lambda \ \beta\ _1$	Performs feature selection, shrinks some coefficients to zero.
ElasticNet Regression	Combines L_1 and L_2 penalties.	$J(\beta) = \sum (y_i - \hat{y}_i)^2 + \lambda_1 \ \beta\ _1 + \lambda_2 \ \beta\ _2^2$	Balances Ridge and Lasso, handles correlated predictors.
K-Nearest Neighbors (KNN)	Prediction based on average of k nearest neighbors.	$\hat{y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$	Simple, non-parametric, sensitive to k and scaling.

AdaBoost Regressor	Sequentially trains weak learners and combines them.	$F(x) = \sum_{m=1}^M \alpha_m h_m(x)$	Focuses on difficult samples, risk of overfitting.
Proposed Model: AffordLearn	Ensemble based on XGBoost and Gradient Boosting with Random Forest as meta-learner Random Forest as meta-learner	$\hat{y} = f_{meta}([x, \hat{y}_{XGB}, \hat{y}_{GBR}])$	Integrates the advantages of a variety of models: input predictions of XGBoost and Gradient Boosting used to generate predictions in the Random Forest meta-model to achieve better generalization, less bias and variance.
Evaluation Metrics			
R^2 Score	Proportion of variance explained.	$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$	Indicates goodness of fit, higher is better.
Root Mean Squared Error (RMSE)	Measures average magnitude.	$RMSE = \sqrt{\frac{1}{N} \sum ((y_i - \hat{y}_i))^2}$	Penalizes large errors, interpretable in original units.
Mean Absolute Error (MAE)	Measures average absolute difference between predictions and actuals.	$MAE = \frac{1}{N} \sum y_i - \hat{y}_i $	Robust to outliers, interpretable.

Models Interpretation

At a realistic decision-support system, interpretability equals in importance to accuracy. To increase the degree of transparency, the methodology also used post-hoc explainability models SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to conduct the analysis of the feature importance and the behavior of the model. SHAP was used to calculate the importance of global features by assigning a value of contribution to each input feature to all the predictions. It also gave an idea of the strength with which expected costs of education depended on variables like tuition fees, annual rent, and living expenses. Conversely, LIME had instance-level

interpretability, which produced locally linear surrogate models around each prediction.

The combination of SHAP and LIME allowed obtaining macro- and micro-level insight into the decision process of the model. By enabling stakeholders students, universities and policymakers to see not only what features were most important, this dual interpretability framework enabled the model to explain the derivation of every prediction.

The inclusion of interpretability techniques has changed the AffordLearn to a predictive instrument to a transparent analytical framework capable of guiding a credible decision-making process in the area of international education cost prediction.

Proposed Model: AffordLearn

The proposed ensemble regression model described in Figure 5, uses Extreme Gradient Boosting (XGBoost) and Gradient Boosting Regressor (GBR) as base learners, and join them with a Random Forest (RF) meta-learner by using a stacking algorithm. The base learners represent the complementary facets of the nonlinear interactions among the features, whereas the meta-learner combines their results to obtain better robustness as well as generalization. Let the dataset be represented as:

$$D = \{(x_i, y_i)\}_{i=1}^N$$

Where $x_i \in \mathbb{R}^d$ represents features vectors and $x_i \in \mathbb{R}$ is the target variable (Total Education Cost).

Base Learners

The first-level models are defined as:

$$\begin{aligned}\hat{y}_i^{(1)} &= fXGB(x_i; \Theta_1) \\ \hat{y}_i^{(2)} &= fXGR(x_i; \Theta_2)\end{aligned}$$

Where $fXGB$ and $fXGR$ denote the XGBoost and Gradient Boosting prediction functions parameterized by Θ_1 and Θ_2 , respectively. Each base learner minimizes the regularized objective:

$$\mathcal{L}_k = \sum_{i=1}^N l(y_i, \hat{y}_i^{(k)}) + \Omega(\Theta_k)$$

Where l . Represents the squared loss and $\Omega(\Theta_k)$ denotes the regularization term to prevent overfitting.

Meta-Learner

The outputs of the base learners form a meta-feature vector:

$$z_i = [\hat{y}_i^{(1)}, \hat{y}_i^{(2)}]$$

which is then used as input for the Random Forest meta-learner:

$$\hat{y}_i = fRF(z_i; \Theta_3)$$

The final prediction function of the ensemble model is:

$$\hat{y}_i = F(x_i) = fRF(fXGB(x_i), fXGR(x_i))$$

The overall optimization goal of the proposed model is to minimize the mean squared error (MSE):

$$\min_{\Theta_1, \Theta_2, \Theta_3} \sum_{i=1}^N (y_i - F(x_i))^2$$

Table 2 describes the metadata and parameters of proposed AffordLearn model.

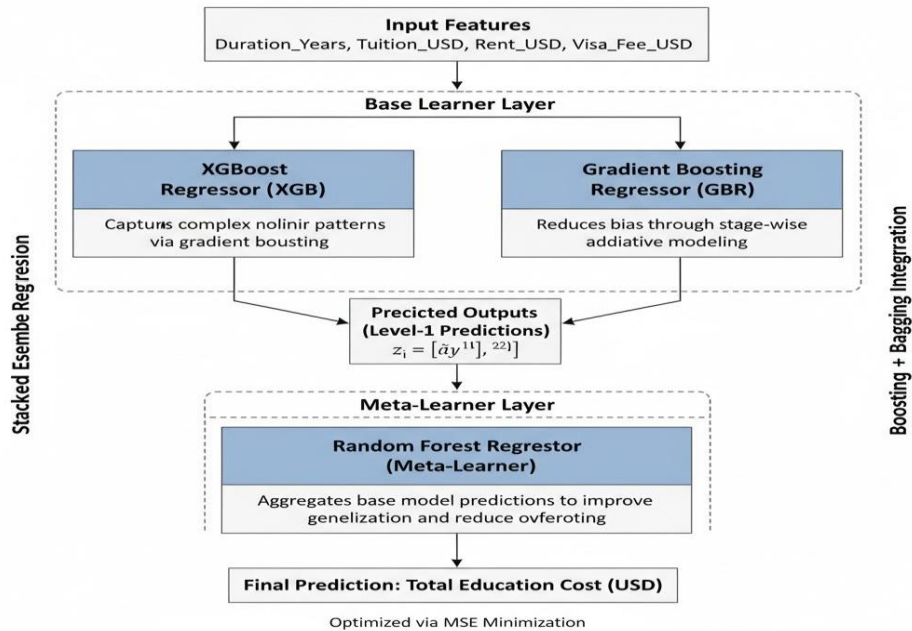


Figure 5. AffordLearn: Proposed Ensemble Regression Model

Table 2. Metadata and Parameters of the Proposed AffordLearn Model

Property	Description / Parameter Values
Model Type	Stacked Ensemble Regressor
Base Learners	XGBoost Regressor and Gradient Boosting Regressor
Meta Learner	Random Forest Regressor
Learning Objective	Minimize Mean Squared Error (MSE)
Boosting Strategy	Additive, gradient-based residual learning
Meta-level Input	Predictions from base learners (XGB, GBR)
Optimization Algorithm	Gradient Descent (boosting), Bagging (meta-level)
Regularization	L1 and L2 (XGB), maximum depth limitation (RF, GBR)
Hyperparameters (XGBoost)	n_estimators = 200, learning_rate = 0.1, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8, reg_alpha = 0.1, reg_lambda = 1.0
Hyperparameters (Gradient Boosting)	n_estimators = 200, learning_rate = 0.1, max_depth = 4, min_samples_split = 4
Hyperparameters (Random Forest)	n_estimators = 200, max_depth = 6, min_samples_split = 4, random_state = 42

Results and Discussion

In order to evaluate the predictive ability of the proposed AffordLearn model, an extensive experimental analysis was carried out, which was applied to various baseline regression methods. These were linear (Linear Regression, Ridge, Lasso

and ElasticNet), tree (Decision Tree, Random Forest, Extra Trees, Gradient Boosting), boosting (XGBoost, AdaBoost), kernel-based (SVR) and instance-based (KNN). Each of the models was trained under the same experimental conditions and with the same preprocessed data and

evaluation metrics to ensure they are fair and comparable.

This paper has compared the performance of thirteen different regression models to determine total cost of international education as shown in Figure 6. The evaluation of the performance was based on four metrics, which were the coefficient of determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and computation time. The findings show that ensemble-based models were significantly more effective than the traditional linear and neural models, and the suggested AffordLearn performed the best in all aspects (the highest accuracy and the lowest values of errors). This shows that the meta-learning approach of integrating multiple ensemble learners into a single prediction model produces a more robust and stable model of education cost prediction.

Table 3 compares the results of the several regression models compared in terms of R^2 , RMSE, MAE, and computation time. Ensemble tree-based models such as Random Forest, Extra Trees and XGBoost all had high predictive accuracy with the R^2 values of over 0.969. Low values of RMSE and MAE were also observed in these models, which proves their high efficiency in the representation of complex and nonlinear

dependencies, among the variables related to cost. Gradient Boosting demonstrated competitive performance ($R^2 \approx 0.958$), but also a bit less than the best ensemble approaches because it is hyperparameter sensitive.

Linear regression, Ridge, Lasso, and Elastic Net linear models showed a moderate level of performance with (R^2 ranging from 0.81 to 0.92) that demonstrates their poor ability to capture higher order interactions between features (tuition, accommodation, and visa costs). Instance-based learners such as K-Nearest Neighbors (KNN) and Decision Tree had lower predictive accuracy but enjoyed the benefit of computing simplicity and so are applicable in lightweight or real time applications where efficiency is of primary concern.

AffordLearn, the proposed stacked ensemble model, recorded the best overall performance with $R^2 = 0.9723$, RMSE = 1.97 and MAE = 1.46. This compares well to current ensemble techniques, and combining multiple learners via a meta-learning framework has been found to improve both prediction accuracy and generalization. Overall, the results show that hybrid ensemble models, especially the proposed AffordLearn model, are the best method to obtain correct and understandable education cost predictions.

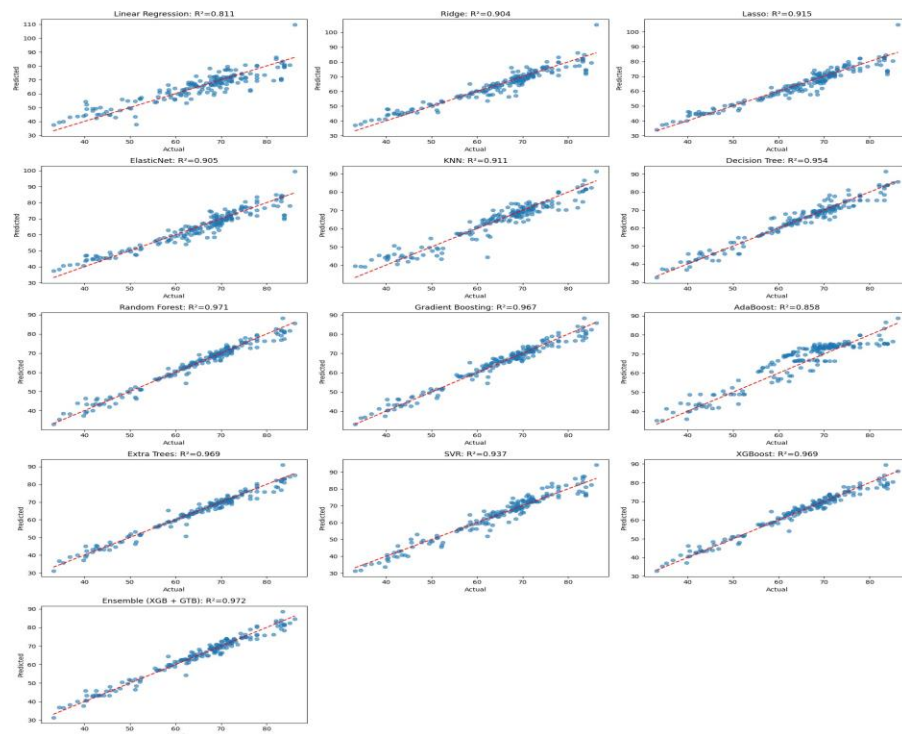


Figure 6. R^2 Results

Table 3. Performance summary of regression models in terms of R^2 , RMSE, MAE, and computation time

Model	R^2	RMSE	MAE	Time (s)
Random Forest	0.9703	2.0451	1.4480	3.0968
Extra Trees	0.9699	2.0571	1.2863	3.4367
XGBoost	0.9693	2.0779	1.3811	0.7458
Gradient Boosting	0.9581	2.4301	1.9262	0.4214
Decision Tree	0.9538	2.5514	1.7022	0.0849
SVR	0.9368	2.9840	2.1608	0.1004
Lasso	0.9155	3.4504	2.3914	0.3618
KNN	0.9107	3.5453	2.5820	0.0658
ElasticNet	0.9047	3.6631	2.6618	0.1728
Ridge	0.9038	3.6810	2.5241	0.0391
AdaBoost	0.8516	4.5711	3.8553	0.4930
Linear Regression	0.8112	5.1567	3.7605	0.0834
Proposed Ensemble Model	0.9723	1.9734	1.4675	13.7113

Conclusion and Future Work

This article has provided an in-depth investigation into machine learning methods to predict the total cost of international education, combining several financial elements including tuition,

accommodation, cost of living, visas and insurance costs. A comprehensive comparison of a wide range of regression algorithms, such as linear, tree-based and ensemble algorithms, based on uniform evaluation measures (R^2 , RMSE, MAE and even

computation time, was systematically used. The results showed that ensemble-based learners are far superior to traditional linear and neural models in terms of predictive accuracy and robustness.

The proposed stacked ensemble model, AffordLearn, achieved the highest overall performance $R^2 = 0.9723$, $RMSE = 1.97$, and $MAE = 1.46$, outperforming all baseline algorithms. AffordLearn also provides a good foundation for information-based decision support systems in the education sector. The fact that it works very well in terms of model transparency makes it a reliable analytical tool in predicting study costs in different countries and programs.

Further studies could be aimed at extending this study in various directions. To enhance temporal flexibility, first, the data may be expanded with dynamic economic indicators (currency exchange rates, inflation, and institutional ranking indices). Second, more sophisticated hybrid deep learning networks like transformer-based regressors or graph neural networks might be considered to advance feature representation and contextual representation further. Third, multi-source datasets (e.g., government data, scholarship databases, and real-time living cost indices) could be integrated to come up with a more comprehensive framework of estimation.

REFERENCES

- [1] B. D. Baker and C. E. Richards, "A comparison of conventional linear regression methods and neural networks for forecasting educational spending," *Economics of Education Review*, vol. 18, no. 4, pp. 405–415, 1999.
- [2] Z. Patel, A. Porwal, K. Bhandare, and J. Woo, "US college net price prediction comparing ML regression models," *arXiv preprint arXiv:2406.08071*, 2024.
- [3] M. Lu, Q. Hou, S. Qin, L. Zhou, D. Hua, X. Wang, and L. Cheng, "A stacking ensemble model of various machine learning models for daily runoff forecasting," *Water*, vol. 15, no. 7, p. 1265, 2023.
- [4] *Security and Communication Networks*, "Retracted: Prediction of higher education cost and analysis of sharing ability based on artificial neural network," 2023.
- [5] R. Kablan, H. A. Miller, S. Suliman, and H. Frieboes, "Evaluation of stacked ensemble model performance to predict clinical outcomes: A COVID-19 study," *International Journal of Medical Informatics*, vol. 175, p. 105090, 2023.
- [6] Y. Ma, X. Tu, X. Luo, L. Hu, and C. Wang, "Machine-learning-based cost prediction models for inpatients with mental disorders in China," *BMC Psychiatry*, vol. 25, no. 1, p. 33, 2025.
- [7] L. Chen, C. Xu, W. H. Lim, A. Sharma, S. S. Tiang, K. S. Chong, E.-S. M. El-Kenawy, A. A. Alhussan, M. E. Eid, and D. S. Khafaga, "Transparent and reliable construction cost prediction using advanced machine learning and explainable AI," *Engineering Science and Technology, an International Journal*, vol. 70, p. 102159, 2025.
- [8] L. Sun, T. Wang, B. Hui, Y. Li, and L. Tian, "Explainable and personalized medical cost prediction based on multitask learning over mobile devices," *Mobile Information Systems*, vol. 2022, p. 8966266, 2022.
- [9] M. Mandolini, L. Manuguerra, M. Sartini, G. M. Lo Presti, and F. Pescatori, "A cost modelling methodology based on machine learning for engineered-to-order products," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108957, 2024.
- [10] T. İnan, T. Narbaev, and Ö. Hazır, "A machine learning study to enhance project cost forecasting," *IFAC PapersOnLine*, vol. 55, no. 10, pp. 3286–3291, 2022.
- [11] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, pp. 1189–1232, 2001.
- [12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

- [13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [14] L. Zhu, W. Lu, C. Luo, Y. Xu, and Z. Wang, "An ensemble optimizer with a stacking ensemble surrogate model for identification of groundwater contamination source," *Journal of Contaminant Hydrology*, vol. 267, p. 104437, 2024.
- [15] M. S. Tahmouresi, M. H. Niksokhan, and A. H. Ehsani, "Enhancing spatial resolution of satellite soil moisture data through stacking ensemble learning techniques," *Scientific Reports*, vol. 14, no. 1, p. 25454, 2024.
- [16] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, "It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy," in *Proc. 2022 ACM Conf. Fairness, Accountability, and Transparency*, 2022, pp. 248–266.
- [17] O. Habib, M. Abouhamad, and A. Bayoumi, "Ensemble learning framework for forecasting construction costs," *Automation in Construction*, vol. 170, p. 105903, 2025.
- [18] N. T. Qassem and I. A. Saleh, "Prediction of software cost estimation using stacking ensemble learning method," *Procedia Computer Science*, vol. 258, pp. 727–736, 2025.
- [19] B. Naderalvojud and T. Hernandez-Boussard, "Improving machine learning with ensemble learning on observational healthcare data," in *Proc. AMIA Annual Symposium*, vol. 2023, p. 521, 2024.
- [20] L. Chen, S. S. Tiang, K. S. Chong, A. Sharma, T. Berghout, and W. H. Lim, "Predicting construction costs with machine learning: A comparative study on ensemble and linear models," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 3, pp. 624–645, 2025.
- [21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [22] M. T. Ribeiro, S. Singh, and C. Guestrin, "“Why should I trust you?”: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [23] R. Alamri and B. Alharbi, "Explainable student performance prediction models: A systematic review," *IEEE Access*, vol. 9, pp. 33132–33143, 2021.
- [24] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.
- [25] C. Lozano-Murcia, F. P. Romero, J. Serrano-Guerrero, and J. A. Olivas, "A comparison between explainable machine learning methods for classification and regression problems in the actuarial context," *Mathematics*, vol. 11, no. 14, p. 3088, 2023.
- [26] A. Papenmeier, G. Englebienne, and C. Seifert, "How model accuracy and explanation fidelity influence user trust," *arXiv preprint arXiv:1907.12652*, 2019.
- [27] D. Slack, S. A. Friedler, C. Scheidegger, and C. Dutta Roy, "Assessing the local interpretability of machine learning models," *arXiv preprint arXiv:1902.03501*, 2019.
- [28] K. Sokol and P. Flach, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," in *Proc. 2020 Conf. Fairness, Accountability, and Transparency*, 2020, pp. 56–67.
- [29] D. Hammann and M. Wouters, "Explainability versus accuracy of machine learning models: The role of task uncertainty and need for interaction with the machine learning model," *European Accounting Review*, pp. 1–34, 2025.

- [30] Kaggle, "International Student Cost Analysis and Prediction." [Online]. Available: [Kaggle dataset](#). [Accessed: Oct. 13, 2025].
- [31] Conversational AI in Healthcare: An Exploration of Large Language Models for Patient Support and Communication. (2025). *International Journal of Advanced Computing & Emerging Technologies*, 1(5), 65-79. <https://doi.org/10.64796/2yv51w38>.
- [32] The Role of Artificial Intelligence in Academic Writing: A Literature Review. (2025). *Annual Methodological Archive Research Review*, 3(11), 242-258. <https://doi.org/10.63075/qk7n8h28>.
- [33] AI-Powered Sentiment Analysis for Social Media Opinion Mining: A Hybrid NLP and Machine Learning Approach. (2025). *International Journal of Advanced Computing & Emerging Technologies*, 1(5), 1-15. <https://doi.org/10.64796/7ytc4b98>.
- [34] Muhammad Usman Javeed, Sammar Fatima, Mirza Mumtaz Zahoor, Muhammad Azhar, Zeeshan Raza, Shafqat Maria Aslam, & Muhammad Nauman. (2025). MindMate: An Emotion-Aware Generative AI System for Personalized Mental Health Support. *Journal of Computing & Biomedical Informatics*, 10(01). <https://doi.org/10.56979/1001/2025/1167>.
- [35] Shakeel, H. ., Akram, M. ., Javeed, M. U., Azhar, M. ., Aslam, S. M. ., Saifullah, & Mumtaz, M. T. . (2025). LncRNAs Disease: A text mining Approach to Find the role of lncRNA in Aging. *Journal of Computing & Biomedical Informatics*, 9(01). Retrieved from <https://jcbi.org/index.php/Main/article/view/1000>.
- [36] Impacts of Cloudburst Events on Biodiversity in Pakistan's Northern Areas and Development of a Machine Learning Model for Cloudburst Prediction. (2025). *International Journal of Advanced Computing & Emerging Technologies*, 1(4), 44-63. <https://doi.org/10.64796/y82wy489>.