ISSN: 2182-2069 / E-ISSN: 2182-2077

OptCatB: Optuna Hyperparameter Optimization Model to Forecast the Educational Proficiency of Immigrant Students based on CatBoost Regression

Selvaprabu Jeganathan[1], Arun Raj Lakshminarayanan[2*], Saravanan Parthasarathy[3], A. Abdul Azeez Khan[4], and K. Javubar Sathick[5]

[1]Research Scholar, Department of Computer Science & Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.
Selva_cse_phd_17@crescent.education, https://orcid.org/0000-0003-0004-3214

[2*]Associate Professor, Department of Computer Science & Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.
arunraj@crescent.education, https://orcid.org/0000-0001-8181-5022

[3]Research Scholar, Department of Computer Science & Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.
saravanan_cse_2019@crescent.education, https://orcid.org/0000-0002-6229-8688

[4]Associate Professor, Department of Computer Applications, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.
abdulazeezkhan@crescent.education, https://orcid.org/0000-0001-6960-752X

[5]Associate Professor, Department of Computer Applications, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.
javubar@crescent.education, https://orcid.org/0000-0002-2248-8380

Received: December 27, 2023; Revised: February 12, 2024; Accepted: March 15, 2024; Published: May 30, 2024

Abstract

A person's poor educational performance or academic success would hinder the struggle against poverty that plagues humanity, particularly for children who are close to completing high school. This study examines the PISA dataset containing immigrant student information from nations like the UAE, New Zealand, Canada, Qatar, Spain, and Australia. As a result, they place a high priority on analyzing the performance of immigrant students to provide them with a high-quality education. Based on the data analysis and interpretation of the findings, factors like early arrival, late arrival, wealth factors, family circumstances, and a multitude of other socioeconomic factors have an influence on the performance of students in reading, math, and science scores. The proposed OptCatB model makes predictions regarding the academic success of immigrant students by applying an optimized CatBoost regressor by keeping reading, math, and science as target variables. We trained the model using the optimized parameters after tuning the hyperparameters of the CatBoost algorithm by using a hyperparameter optimization technique termed Optuna. The OptCatB model outperformed compared to the other selected regression models with RMSE of 54.231, MAE of 43.104, MAPE of 9.931 and RSE of 0.54.

Journal of Internet Services and Information Security (JISIS), volume: 14, number: 2 (May), pp. 111-132.
DOI: *10.58346/JISIS.2024.I2.008*

*Corresponding author: Associate Professor, Department of Computer Science & Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.

Keywords: Regression Models, CatBoost Regressor, Educational Data Mining, PISA, Machine Learning, Optuna, Hyperparameter Optimization, Immigrant Students.

1 Introduction

Through education, individuals learn the skills necessary to perceive, interpret, and function in their current environment, as well as successfully adapt to rapidly changing circumstances. Almost everyone acknowledges that education has the effects of informing, training, and equipping, but much depends on the number of years spent in school and the quality of education acquired. It is uncertain how many years of education are required, as well as whether there is a minimum or maximum number of years required. The link between years of schooling and acquired knowledge varies from country-to-country, dependent on variables such as teacher preparation, educational resources, and school year length. Literacy is measured using standardized levels of education such as None, Elementary, Secondary, and Higher, as well as the number of years spent in school. Even though classification by level of education doesn't change between levels attended and levels finished, this difference needs extra thought.

Education is meant to promote inventive behavior and cognition, not just literacy. Educational conditioning is progressive but restricted education is less likely to give these benefits. Today, most nations' education systems must ensure that immigrant students receive the academic skills needed for successful resettlement. Large representative samples of students who have taken the same tests of academic knowledge and skills would be excellent for study in order to better understand the factors that affect immigrant students' academic abilities (Akther & Robinson, 2014). According to researchers, the achievement test score disparity for immigrant students has been linked to a number of potential factors. A research study involving more than 10 countries, including Slovenia, Switzerland, Hong Kong, Argentina, Turkey, and Costa Rica, found that a number of individual and family level attributes, such as lower income, were linked to immigrant students' performance on achievement tests. The immigrant achievement gap and the question of whether it occurs beyond national boundaries have been the subject of certain studies on immigrant education. Most studies compare immigrant and native student academic performance by accounting for gender, home language, books obtained by the household, parents' profession and other indicators of poverty. Various theories, including Human Capital Theory (Ross, 2023), Social Mobility Theory (Team, 2023), and Signaling Theory (Libretexts, 2021), discuss the relationship between education and poverty. Additional research emphasizes the need for policymakers to focus on improving the quality of education to reduce poverty vulnerability and enable students to compete in the global labor market (Cameron, 2012; Inoue et al., 2023; Lewin, 2023, Ngepah et al., 2023).

The tests that are most frequently used in the research on immigrant schooling vary in a number of ways. PISA evaluates the reading, arithmetic, and science abilities of 15-year-old children. PISA examines students rather than being connected to the curriculum in schools by assessing "how well students are prepared for active participation after graduation and how well they can apply their knowledge." (Andon et al., 2014). Australia's involvement in the worldwide PISA surveys and the implementation of high-stakes performance exams like NAPLAN haven't done much to address the actual educational needs of immigrant populations. Although first-generation immigrants performed better than both local and foreign-born students in PISA 2015 and students from households where at least one parent was born abroad performed better in NAPLAN (National Assessment Program – Literacy and Numeracy), these generalized results are not especially helpful. According to the research

above, some ethnic populations in Australia's educational system do far better than others, a result that the Australian report's reluctance to separate the PISA immigrant data actually obscures (Welch, 2018).

Although there were some increases in the PISA Assessment, the performance of Pasifika learners, a specific group of immigrants in New Zealand, has remained subpar in comparison to other immigrant communities. The Pasifika Education Plan (2013–2017) has been drafted by the government to increase the academic performance of the Pasifika community, to increase participation in early childhood education, increase national achievement and decrease school dropouts. The Pasifika Competency Framework was made to help teachers do a better job of guiding their students (Poskitt, 2018).

Browning & Rigolon, (2019) examined the benefits of green space on academic performance by also linking socioeconomic factors. Most data on writing exam scores were non-significant, while the correlation between school green infrastructure and the effects of socioeconomic status, gender, and urbanization on moderation produced conflicting results. Cortes Pascual, et al., (2019) correlates the gender of the student to study the effects of executive functions such as Working Memory and flexible thinking on academic achievement. Physiological and neurological factors cause female students to grow more rapidly than male children during primary school. Based on evidence from various socioeconomic levels, the education system in rural versus urban areas has an indirect effect on intellectual development, which is proportionate to academic success.

The remaining sections of the paper are structured as follows: The section two provides a summary of recent advancements in educational data mining research. The characteristics and pre-processing of the data collection are detailed in Section 3. The fourth section presents a summary of the results of exploratory data analysis. Section 5 describes the outcome of the experiment. The sixth section concludes by emphasizing on the study's findings and future research.

2 Related Work

Scores earned by the student in reading, science and mathematics could be used to predict his or her academic performance using the PISA dataset. Factors associated with higher education and questions posed during data collection contribute to the wide range of educational variables. Young & Caballero, (2021) conducted a study on an educational dataset using cutting-edge machine learning methods. The Log-F and Firth penalization strategies have been implemented to improve the algorithmic results, which varied due to imbalance of data. Suggested using a penalized logistic regression model for smaller datasets and a Random Forest or conditional inference forest algorithm for datasets with more than 1,000 instances, which he refers to as medium or bigger datasets (Jonnerby et al., 2023; Srinivasa Rao et al., 2023). For data mining and classification CNN plays a crucial role in various fields (Camgozlu, & Kutlu, 2023).

In educational data mining, data validation is a critical step in the process. Feldman-Maggor et al., (2021) concentrates on data preparation that could be applied to educational data mining for future research. The study was organized by gathering data from Israel Open University and Weizmann Institute of Science, and the acquired data was evaluated by designating a test user and creating a log file. Having constructed a database by comparing the data log file with the test log file, the database's data has been cleaned, filtered, and aggregated in preparation for educational data mining. Figure 1 depicts the data mining pipeline in education.

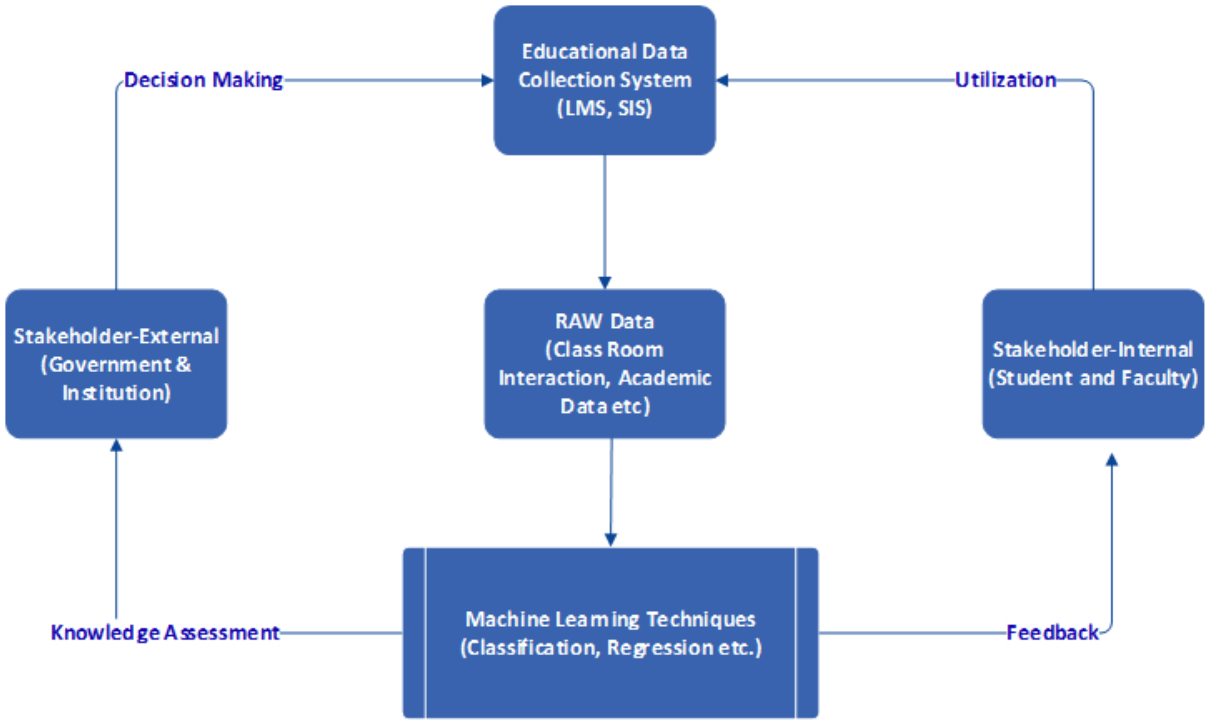

Figure 1: Application of data mining in education

Educational Institutions must develop more accurate strategies for predicting student withdrawals. Using Analytic Examination, Classic Machine Learning, and Deep Learning methodologies, (Prenkaj, 2020) conducted an analysis on student dropout. Classic techniques include rule-based models, ensemble models, Nave Bayes, and Support vector machines, while the deep learning model includes recurrent and convolutional neural networks (Juma et al., 2023).

Recently, pedagogy methods have evolved from traditional classroom learning to online learning, which affects students' procrastination and may eventually lead to drop-out. The rise of this behavior trend among students could be predicted by analyzing their assignment-submission patterns. The linear support vector machine technique identifies students' procrastination by linking assignment grades and submission dates with an accuracy of 96% (Hooshyar, 2020). To lower student dropout rates, (Berens et al., 2019) offer a system for early detection based on data collected from State and Private universities of applied sciences in Germany, the dataset includes demographic, previous, and current academic data. Logit Boost, Neural Network, Bagging Random Forest, and AdaBoost algorithms have forecasted the dropout prediction. At the end of the fourth semester, the prediction accuracy and AUC percentage improved to 90% for the AdaBoost method, which was superior to the other three (Bharathi & Rekha, 2023).

Rebai, et al., (2020) Utilized a two-stage study on a dataset from a Tunisian secondary school to determine the important characteristics that influence the academic achievement of students. In the first phase of Data Envelopment Analysis, undesired data outputs are dealt with using the Directional Distance Function Approach. In the second stage, machine learning techniques are used to analyze the elements that influence the academic performance of schools. Regression trees identified factors such as School Size, Class Size, and Parental Pressure as having a substantial impact on academic success. The findings of the Random Forests algorithm indicate that the proportion of females in the school and

the size of the school have the biggest effect on the accuracy of the model's predictions, and it may have a higher effect on school efficiency.

Rastrollo-Guerrero et al., (2020) conducted a literature study on machine learning algorithms for predicting students' academic achievement. In the study of the educational dataset, Decision Trees, Nave Bayes, Support Vector Machine, and Random Forest did well. Using the data acquired from Microsoft's showcase school, the 10-fold cross-validation technique identifies the classifier with the best performance. Using Self-Training, an unlabeled data set is partitioned, which enhances the performance of the classification algorithms in use (Livieris et al., 2019). (Khan & Ghosh, 2021) conducted a literature review on educational data mining articles and found that identifying the prediction variable is the most important effort in an educational dataset, with just 33% of studies able to predict students' performance before the start of the course. In addition, it appears that forecasting success or failure is more prevalent than predicting an overall grade or mark (Liloja & Ranjana, 2023).

Ramaswami et al., (2022) proposed a generic model of academic achievement regardless of the course or its variations. A variety of methods have been tested for their ability to predict binary and multiclass classifications utilizing Random Forest, Naive Bayes, Logistic Regression, KNN and CatBoost. Feature Importance was calculated using the Shapely Additive Explanations approach, which predicted the influence of changing feature values on the prediction. CatBoost is superior to other algorithms in its ability to manage category elements, missing data, and general dataset changes. Beaulac & Rosenthal, (2019). Experimented with a big dataset containing 10-year data on undergraduate students from a Canadian university by building two Random Forest-based classifiers. On the two classification models developed, one predicted the likelihood that the student would successfully complete the course, while the other predicted the department of students who successfully completed the course.

Coleman et al., (2019) worked on identifying students at risk in new United States school districts. XGBoost, Support Vector Machine, Logistic Regression, Random Forest, and Decision Trees were used to generate the initial model. Random Forest outperformed other algorithms with an estimator value of 15 and a maximum depth value of 10 when applied to missing variables. The model achieved a 93% accuracy in forecasting the dropout of 9th through 12^{th} grade students in current school districts; however, its performance declined dramatically to 81% when the number of years of student data was raised to 11 years. The accuracy of the model was determined to be 81% when evaluated against new school districts.

Costa-Mendes et al., (2021) analyzed the bias in machine learning models for predicting student grades by adopting Random Forest, Support Vector Machine, and extreme boosting machine approaches with the Portuguese educational ministry's educational dataset. The outcomes demonstrated that the methodological bias has little effect because the algorithms work identically throughout the study.

Levy et al., (2020) compared traditional statistical techniques such as Linear Regression and multilevel modelling to machine learning techniques. This experiment includes data from 3,026 students from 153 schools that participated in the Luxembourg School Monitoring Program's grade 1 and grade 3 standardized achievement tests. Effectively measuring academic progress using machine learning techniques while maintaining classical models as a foundation. The performance is evaluated by analyzing Math and Language achievement results. Linear Boosting, neural networks, and Random Forest were the most effective machine learning models when compared to the other multilevel models tested. According to (Levy et al., 2020) multilevel models have a high predictive accuracy on the dataset, but will not be applicable to other data contexts. Hussain & Khan, (2023) proposed a hybrid model which utilizes regression techniques to forecast the scores and utilized classification techniques to identify the accuracy by grouping the students based on historic academic records only by using the academic scores

and the socio-economic factors are neglected. Yağcı (2022) forecasts the final exam grades of undergrads using their interim test scores as the source data. Among the machine learning techniques, Logistic Regression, Naive Bayes, and k-nearest neighbor were evaluated in order to predict the students' final exam grades, the results showed a 70-75% accuracy based on the Turkish dataset which comprises only 1854 records (Srinivas & Katarya, 2022). The recent technology can link the entire globe while improving the communication from various locations (Danh, 2020).

Despite the extensive exploration of machine learning and data mining techniques in predicting academic achievement, there exists a notable research gap pertaining to the incorporation of socio-economic attributes into these models. The current literature predominantly focuses on utilizing students' performance records without sufficiently considering the nuanced impact of socio-economic factors such as parental education, income levels, and neighborhood characteristics. This gap poses a limitation in understanding the holistic interplay between socio-economic attributes and traditional academic indicators in predictive models.

Addressing this research gap is crucial for developing a more comprehensive and nuanced understanding of the factors influencing academic success. By integrating socio-economic attributes into predictive models, future research has the potential to enhance the accuracy of predictions and contribute valuable insights for educational institutions, policymakers, and practitioners seeking to implement targeted interventions and support systems. The proposed research direction involves the development and validation of predictive models that encompass both socio-economic attributes and traditional academic indicators, fostering a more inclusive and equitable approach to educational data mining and machine learning. This comprehensive exploration could significantly contribute to advancing knowledge in the field and improving educational outcomes for diverse student populations.

3 Materials

1) Data Set

The Programme for International Student Assessment is conducted by OECD (Organization for Economic Co-operation and Development) to measure 15-year-old student ability in Maths, Science and Reading. OECD studies the 65 nations that account for 90% of global economic activity, Student proficiency in math and science is a reliable predictor of the future economic health of the nation. Individual student, school administrator, and parent responses are included in the PISA database. In order to carry out their own study of the PISA data, statisticians and professional researchers use the data. In our research, we used the PISA 2018 assessment data to forecast the educational achievement of immigrant students in reading, science, and maths. The data quality was ensured by putting in place training criteria, data processing guidelines, and data integrity checks. The dataset had 612,004 rows and 1119 attributes. Out of the 1119 attributes, 35 are identified to figure out the performance level of students in Maths, Science and Reading. The top 5 countries that accept immigrant students are identified by taking the top 5 data count of students based on the immigration status defined in the attribute Index Immigration Status which contains 29894 rows.

According to Immigration by Country 2024 (Immigration by Country 2024, n.d.), these countries have the highest immigrant rates: Spain (6.8M), Qatar (2.2M), Canada (8M), Australia (7.7M), and UAE (8.7M). The missing values in the dataset have been pre-processed by using statistical methods. Unanswered Boolean values of the attributes ST011Q03TA, ST011Q04TA and SCHSIZE have been updated with negative values. Missing values in the following attributes have been updated with Most Occurred value attributes include - ST102Q01TA, ST161Q07HA, ST059Q02TA, ST059Q03TA,

IC152Q02HA, SCHLTYPE, MISCED, FISCED, SC017Q08NA, SC001Q01TA, SC053Q12IA, SC150Q05IA, SC164Q01HA, SC152Q01HA and IC152Q03HA. Missing values in the attributes ST097Q01TA, ST104Q04NA, ST161Q03HA, IC008Q02TA, MMINS, LMINS, SMINS and CLSIZE are updated using their Mean Values. Table 1 depicts the characteristics identified for further examination.

Table 1: Attribute details of the dataset

Attribute	Description
CNTRYID	Identifier for Country
CNTSCHID	Identifier for School
IMMIG	Immigration status indicator
ST004D01T	Gender of Students
ST011Q03TA	At your house: a secluded spot for study
ST011Q04TA	Having access to a home computer that can be used for academic purposes
ST013Q01TA	The number of books in your house.
ST097Q01TA	How frequently: students neglect the instructions provided by instructor during class.
ST102Q01TA	How frequently in class: The instructor emphasizes specific learning objectives.
ST104Q04NA	How often does teacher provide me with suggestions on how to do better in class?
ST161Q03HA	I am able to read well.
ST161Q07HA	I need to study something multiple times before I can fully comprehend it.
ST059Q02TA	Number of Maths Classes Per Week That Are Usually Required to Be Attended
ST059Q03TA	Number of Science Classes Per Week That Are Usually Required to Be Attended
IC152Q02HA	Education-related digital media consumption within the past month: Mathematics
IC152Q03HA	Education-related digital media consumption within the past month: Science
IC008Q02TA	Utilization of digital devices for activities outside of the classroom, such as team-based online gaming
PV1MATH	Credible Score 1 in Mathematics
PV1READ	Credible Score 1 in Reading
PV1SCIE	Credible Score 1 in Science
MMINS	Studying timeframe (minutes per week) - Mathematics
LMINS	Studying timeframe (minutes per week) - Test Language
SMINS	Studying timeframe (minutes per week)- Science
REPEAT	Repeatedly attending the same class multiple times
MISCED	Education of Mother based on ISCED
FISCED	Education of Father based on ISCED
SC017Q08NA	Impact of School Infrastructure on teaching
SC001Q01TA	Location of School (Rural, Urban)
SC053Q12IA	Does the School Offer a Reading Group to Encourage Students to Read?
SC150Q05IA	Policies of the school that prioritize equity include reducing class sizes in order to better meet the specific needs of these students.
SC164Q01HA	In the most recent academic year, what percentage of students in final class left school without a certificate?
SC152Q01HA	Is your school providing extra test language sessions outside of regular school hours?
CLSIZE	Size of the Class Room
SCHSIZE	Number of Students in a School
SCHLTYPE	Possession of a School

2) Proposed Methodology

The proposed method employs a mass data set of student assessments and evaluates it using a multitude of regression models by integrating the Optuna hyperparameter optimization technique by conducting the research in two phases using Python as an interface.

Phase I: Figure 2 depicts the exploratory analysis of the academic performance of the students correlated with infrastructure and teaching characteristics. The collected information was graphically represented using visualization tools.

Phase II: Figure 2 depicts the proposed method through which machine learning algorithms are employed to forecast the academic performance of immigrant students. Regression techniques were used to predict academic performance, and evaluation metrics were used to measure how well the machine learning algorithms worked. The proposed expert model OptCatB which is the integration of CatBoost regressor and Optuna hyperparameter modelling technique outperformed other machine learning algorithms in terms of generated metrics. The outcome of the research has been discussed in the Experiments and Results section briefly.

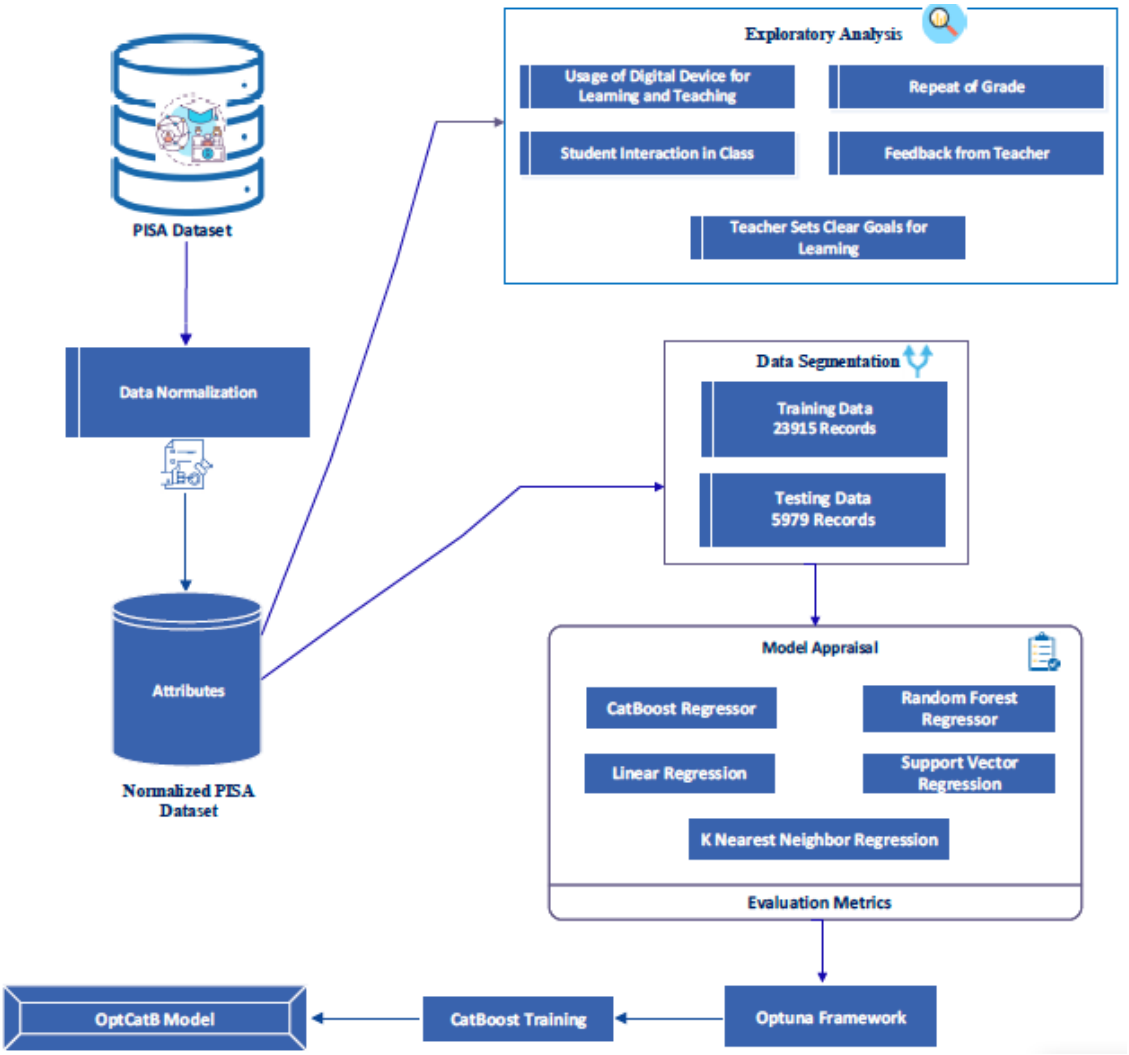

Figure 2: Proposed Methodology

3) Descriptive Analysis

Exploratory analysis helps in decoding the characteristics of data using visualizations. The relationship between attributes, data distribution, outliers, trends, and patterns could be evaluated by analyzing the graphical representations. Simultaneously, it is advantageous to discover appropriate prediction methods. The following visualizations show the basic school-level factors which might influence the academic performance of a student based on data from the PISA assessment.

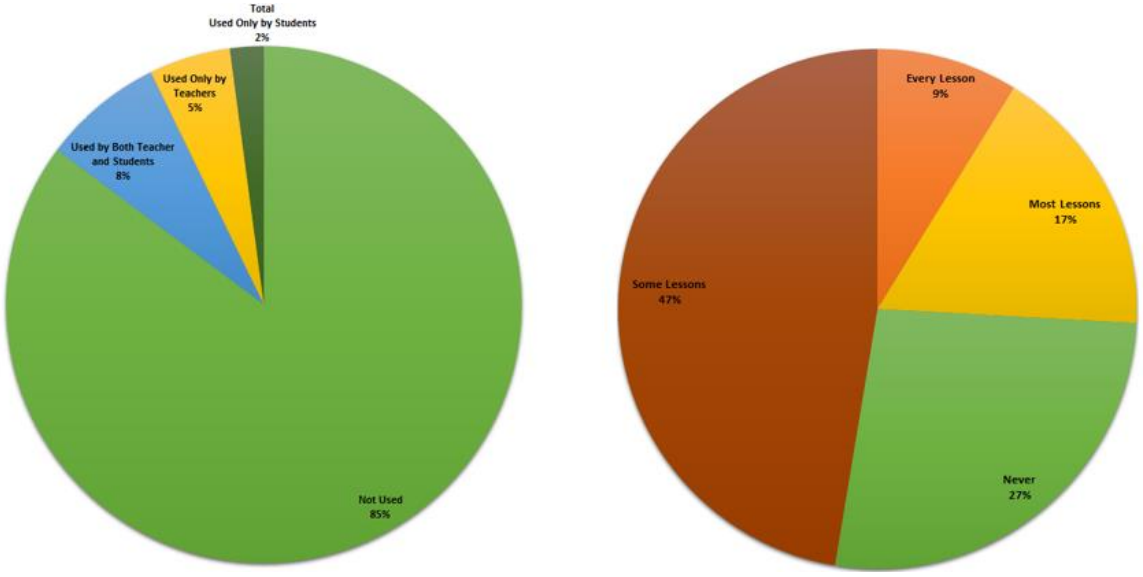

Figure 3: Usage of Digital Devices for Learning and Listening Ratio of Students in Classroom

Figure 3 depicts that digital devices are not used in schools by teachers and students for learning and teaching. Only 8% of students and teachers use digital devices for teaching as well as learning, which indicates that most schools are following only the whiteboard pedagogy technique. Fig. 3 illustrates how interactive the students are in the classroom. Only 9% of the total students are interactive and listening to the lessons taught in the classroom. According to data, 27% of total students are not listening to at least one lesson, which has a direct impact on their academic performance. Also, the highest population of 47% indicates that the ratio of students performing average in their studies is higher in general.

A teacher could provide feedback to students on their performance in lessons taught by having an internal assessment or by having an interactive session with respective students. Fig. 4 depicts whether the teacher provides feedback to the student on their performance in the lesson they taught. Only 18% of teachers provide their feedback to the students on all the lessons they have taught, and 32% of the teachers never provide any feedback to the students on the lessons they teach. Giving students frequent feedback will help them improve their academic performance. Figure 4 depicts whether the teacher sets clear goals for the student's learning; 60% of teachers set clear goals for the student for all lessons, and 24% set goals for most of the lessons they teach, which will help students have a clear vision on every lesson and help them perform better in their academics. The aforementioned teacher-student attributes have a significant impact on the academic achievement of students. In light of the aforementioned variables, it is evident that students with an average performance are more numerous than those who perform well or poorly. Figure 5 shows the proportion of students who repeat the same grade multiple times.

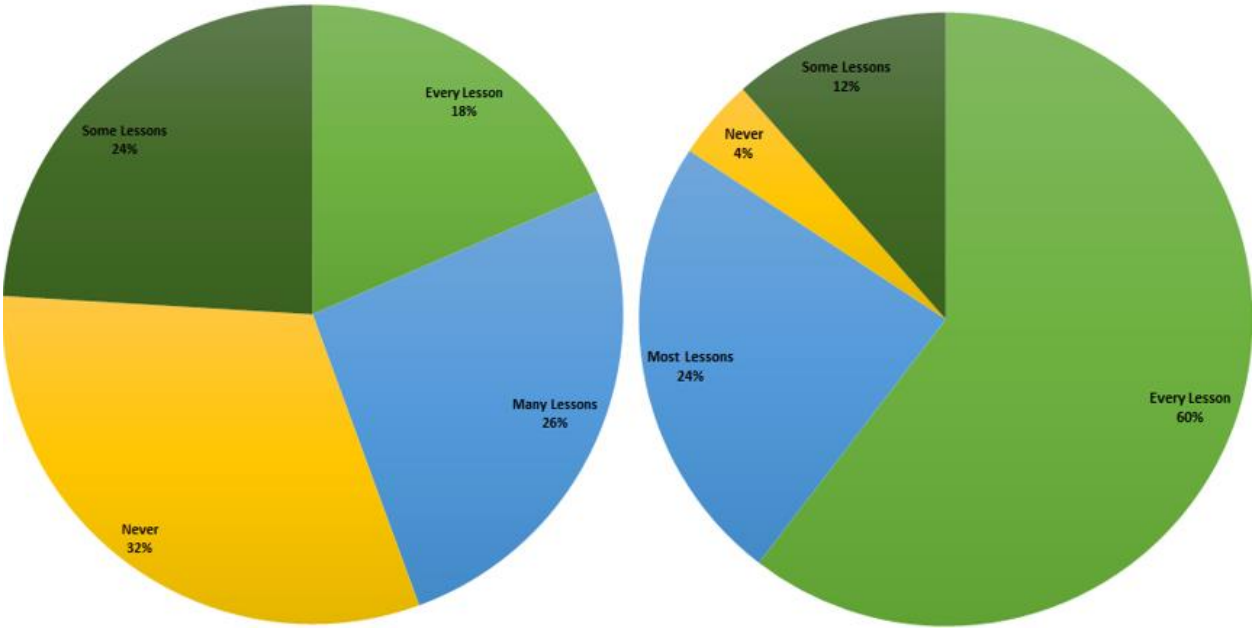

Figure 4: Feedback from Teacher and Goal Setting

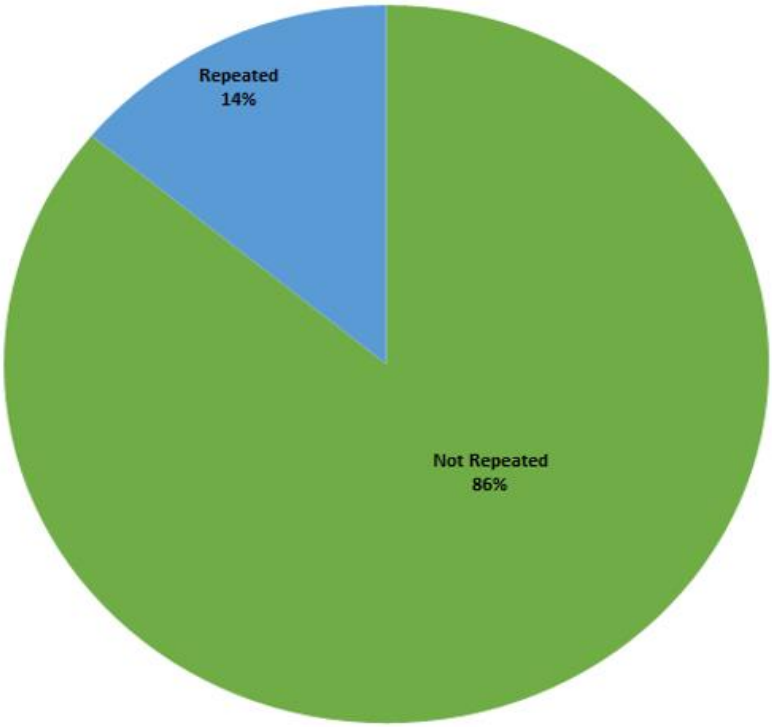

Figure 5: Repeat of Grade

4 Experiments and Results

The cleansed data set comprises information regarding the academic details of the immigrant and the characteristics that are correlated with it. The dataset contains 29894 samples collected from PISA 2018 assessment. The top five nations that accept immigrant students are determined by taking the count of

students with the top five Immigration status values from the Index Immigration Status property. The modelling dataset consists of 80% training data with a row count of 23915 and 20% testing data with a row count of 5979. The target variable in the context of regression models is a real-valued variable Scores obtained in PV1Math, PV1READ and PV1SCIE are identified as the target variables. CatBoost Regressor, Random Forest Regressor, Linear Regression, Support Vector Regression, and K Nearest Neighbor Regressor are the regression algorithms utilized in the data model. CatBoost regressor outperformed all other regression techniques in terms of MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Square Error), R-Square, and MAE (Mean Absolute Error). As stated, the variables are diversified; firstly, PV1Science and PV1Read are measured while PV1Math functions as the objective parameter. The procedure is repeated by measuring PV1Science and PV1Math by keeping PV1Read as the target variable further PV1Math and PV1Read are assessed by keeping PV1Science as the target variable. CatBoost regressor outperformed compared to the other selected regression models in terms of Loss Function (RMSE), MAPE, MAE and Coefficient of determination (RSE). Therefore, in order to improve the effectiveness of the CatBoost regression technique, we have optimized the algorithm by integrating it with Optuna which is a hyperparameter optimization technique. CatBoost successively combines a large number of robust models and develops an accurate forecasting model by greedy searching. When using gradient boosting, the decision trees are fitted one after the other; as a result, the newly fitting trees will learn from the errors committed by the trees that came before them. Until the selected loss function is no longer minimized, the procedure of adding a new function to those that are already there will continue. The learning speed of the CatBoost regressor is higher compared to XGBoost and Light Gradient Boosting regression so XGBoost and Light GBM is not included in the model which we have constructed.

Optimal performance could be obtained in any system with proper tuning of its hyperparameters. Finding a good range for the hyperparameters will have a major effect on the forecasting model. Manual search, random search, grid search, and Optuna are a few of the methods that have been developed to address hyperparameter tuning. Both random and grid search techniques require time to evaluate unnecessary and non-committal search space; hence, they consume a significant amount of time, and grid and random search techniques scarcely learn from the prior refinements. Consequently, we have determined to utilize Optuna as our technique for hyperparameter tuning, Optuna could make a good set of parameters because it constantly learns from previous optimizations and always uses the data from those optimizations (Srinivas & Katarya, 2022).

The tree-Structured Parzen Estimator sampling technique is used for independent parameter sampling, relational parameter sampling is performed by employing Covariance Matrix Adaptation which identifies the correlation between the parameters and search space pruning is accomplished with the Asynchronous Successive Halving technique. Hyperparameter tuning is framed by Optuna as the procedure of maximizing or minimizing an objective function provided a set of hyperparameters and then returning the validation score for that objective function. Optuna builds the objective function incrementally through its association with the trail object. During the execution of the objective function the trail object's method will generate the search space dynamically and using the current run's data, calculates the value of the subsequent hyperparameter to be evaluated and “Should Prune API” is responsible to remove the unwanted trails (Zhou et al., 2022).

Optuna could determine the area in which hyperparameter comparison is most likely to occur and then perform a hyperparameter search in this area based on the data that it has collected. When the new results are received from the execution then the area is updated, and the search has been continued. This process of search is repeated until the optimized hyperparameters which yield better performance is

reached. CatBoost model is trained using the tuned parameters optimized by Optuna, following are the best parameters identified: loss_function, learning_rate, l2_leaf_reg, colsample_bylevel, depth, boosting_type, bootstrap_type, min_data_in_leaf and one_hot_max_size. RMSE value of training data is the loss function, which is measured based on the objective function chosen, Optuna repeats the execution until the maximum optimized value of the loss function is reached. Figure 6 illustrates the schematic architecture of Optuna and Table 2 depicts the associated values of best fit parameters identified. The following are the actual operations carried out by the proposed model – OptCatB, the proposed model uses both boosting modes ordered and plain which is identified by the Optuna framework based on the target variable during runtime. Algorithm 1 provides the corresponding pseudocode.

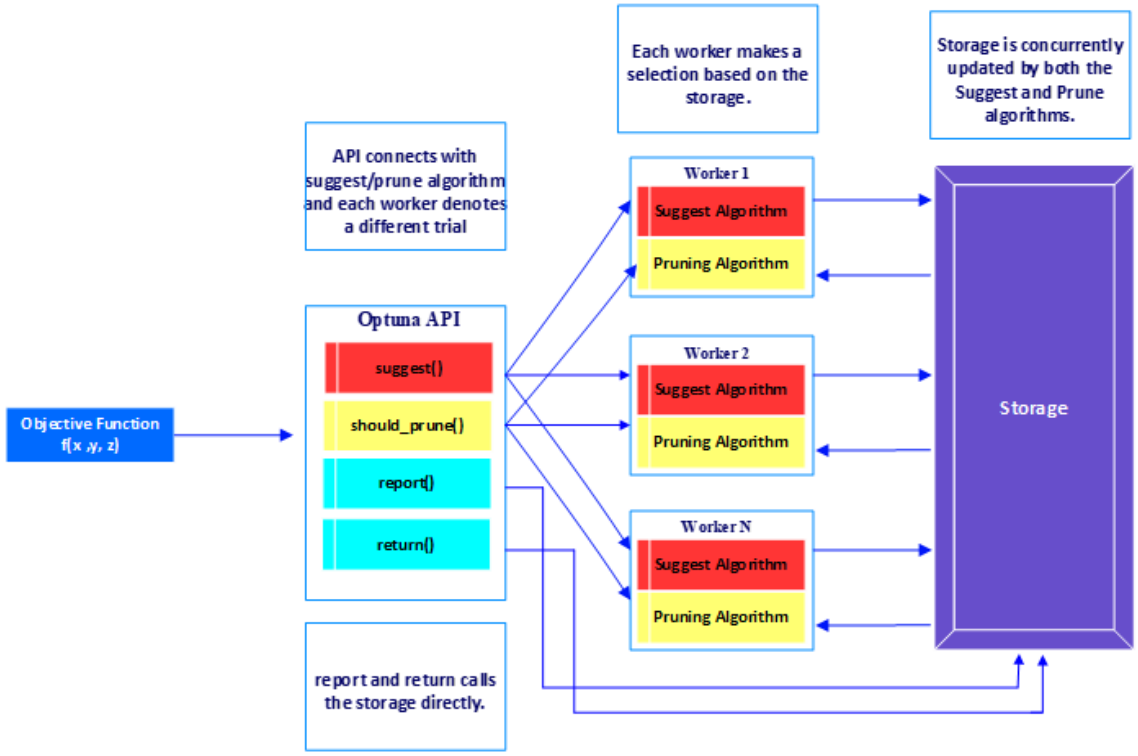

Figure 6: Schematic Depiction of the Architecture of Optuna

Table 2: Best fit parameters Identified from Optuna

Best Fit Parameters	Maths as Target Variable	Science as Target Variable	Reading as Target Variable
loss_function	RMSE	RMSE	RMSE
learning_rate	0.053439159	0.034666773	0.042464631
l2_leaf_reg	0.064584106	0.977688218	0.046104994
colsample_bylevel	0.091365908	0.083427046	0.083148424
Depth	6	9	9
boosting_type	Plain	Plain	Ordered
bootstrap_type	MVS	MVS	Bernoulli
min_data_in_leaf	12	6	18
one_hot_max_size	10	14	16

Algorithm 1: Pseudocode of proposed OptCatB Model

$Input$: $PISA\ academic\ dataset\ and\ hyperparameters$
1. $Read\ the\ dataset$
2. $Split\ dataset\ as\ Training\ and\ testing\ data$
3. $Initialize\ the\ model, \{(X_i, y_i)\}_{i=1}^{n}, I, \alpha, \mathrm{L}, \mathrm{st}, \mathrm{mode}$
4. $Generate\ Independent\ random\ permutation(st+1)\ of\ training\ dataset$
5. $If\ mode = plain\ then$
$Calculate\ the\ leaf\ node\ based\ on\ Target\ Statistic.$
$Mod_{rv}(i) \leftarrow 0\ for\ rv = 1..st\ where\ i: \sigma_r(i) \leq 2^{j+1}$
$Where\ Mod - Model, rv - residual\ value, st -$
$random\ permutation\ \sigma_r\ serves\ in\ choosing\ leaf\ value. To\ produce\ new\ training\ sets, permute\ rows\ of\ ST\ at\ random.$
$Permutation\ process\ is\ repeated\ st\ times\ to\ generate\ new\ unique\ training\ sets$
6. $If\ mode = Ordered\ then$
$For\ each\ j^{th} instance\ in\ data\ samples:$
$Initialize\ Matrix\ Mod(rv, i) = 0$
$rv = 1,2,,,st\ (s\ permutation\ set\ exists)$
$i = 1,2,,,n\ (instance\ ascending\ index\ in\ set\ STrv)$
$Mod(rv, i) denotes\ the\ prediction\ value\ for\ instance\ i\ on\ training\ set\ STrv$
$Mod_{rv,j}(i) \leftarrow 0\ for\ rv = 1..st\ where\ i = 1..2^{j+1}$
7. $For\ each\ permutation\ from\ training\ set\ ST_{rv}$:
7.1. $Employ\ Ordered\ Target\ Statistics\ (TS)\ Encoding\ across\ all\ categorical\ attributes.$

$$X_i^k = \frac{\sum_{Xj \in D_k} \left\{x_j^i = x_k^i\right\} * y_j + ap}{\sum_{Xj \in D_k} \left\{x_j^i = x_k^i\right\} + a}$$

$D_k\ is\ the\ instance\ from\ x_{1\ to}\ x_{(k-1)}$
$a\ is\ weighted\ parameter, P\ is\ target\ avergage\ value$
7.2. $Generate\ new\ ordered\ boosting\ tree: OT$
7.3. $Applying\ new\ tree\ T\ to\ forecast\ each\ permutation\ dataset\ ST_1..ST_s\ and\ update\ Mod$
$Mod_{rv,i} = M_{(rv,i)} - \sigma * OT_i$
$rv = 1,2 \ldots . ST(for\ each\ permuation\ set)$
$i = 1,2 \ldots . n(for\ each\ instance\ in\ set\ ST_r)$
$Repeat\ Step\ 7.1\ to\ 7.3\ to\ build\ I\ trees$
7.4. $Measure\ the\ learned\ model\ by\ evaluating\ the\ loss\ function$
7.5. $Evaluate\ the\ base\ learners$
7.6. $Update\ the\ model\ f(x) = \Sigma_{t=1}^{i} \sum_j \alpha b_j^t \{getLeaf(x, OT_t, applyMode) = j\}$
8. $End\ For$
9. $Result: f(x)$
10. $Create\ the\ object\ for\ Optuna\ API$
11. $Define\ Objective\ Function(trial)$:
11.1. $Load\ Data(training_{data}, validation_{data})$:
11.2. $Define\ param$
$param = \{objective, colsample\_level, depth, boosting\_type, bootstrap, loss\_function,$
$learning\_ratel2\_leaf\_reg, min\_data\_in\_leaf, one\_hot\_max\_size, depth\}$:
11.3. $Define\ bagging\_temperature\ and\ subsample\ based\ on\ bootstrap$
11.4. $regressor = CatBoostRegressor(** param)$
11.5. $regressor(X_{train}, y_{train}, eval_{set} = [(X_{eval}, y_{eval})])$
11.6. $loss = mean\_square\_error(y_{valid}, regressor(X_{valid}))$

11.7. $return\ loss$
11.8. $study = optuna.create\_study(f'catboost - seed\{random\_seed\}')$
11.9. $Optimize\ the\ parameters$
$study.optimize(objective\_function, trials)$
$study.best\_params$
11.10. $return\ (study.best\_params)$
12. $End\_OptCatB\ Algorithm$

The performance of algorithms is summarized in Table 3, Table 4 and Table 5 which focuses Maths, Science and Reading as the target variables respectively.

Table 3: Measuring the Effectiveness of Proposed Model using Maths as Objective Variable

Regressors	RMSE	MAE	MAPE	R^2
Proposed(OptCatB)	54.231	43.104	9.931	0.54
CatBoost	75.321	59.867	13.793	0.42
XG Boost	77.115	61.231	14.068	0.39
Random Forest	79.881	63.688	14.781	0.35
Linear Regression	83.906	66.890	15.545	0.29
SVM Regressor	84.319	67.163	15.634	0.28
KNN Regressor	86.872	69.097	16.129	0.24

Table 4: Measuring the Effectiveness of Proposed Model using Science as Objective Variable

Regressors	RMSE	MAE	MAPE	R^2
Proposed(OptCatB)	54.851	43.471	9.916	0.55
CatBoost	75.138	59.550	13.584	0.43
XG Boost	76.930	61.082	13.882	0.40
Random Forest	79.363	63.005	14.466	0.37
Linear Regression	83.741	66.834	15.406	0.30
SVM Regressor	83.991	66.929	15.487	0.29
KNN Regressor	87.693	70.154	16.245	0.23

Table 5: Measuring the Effectiveness of Proposed Model using Reading as Objective Variable

Regressors	RMSE	MAE	MAPE	R^2
Proposed(OptCatB)	56.040	44.394	10.284	0.60
CatBoost	78.929	62.528	14.485	0.46
XG Boost	81.025	63.922	14.787	0.43
Random Forest	84.152	66.628	15.556	0.39
Linear Regression	88.819	70.940	16.703	0.33
SVM Regressor	89.267	71.237	16.869	0.32
KNN Regressor	91.825	73.003	17.227	0.28

The definition of accuracy is "out of all the predictions generated by our model, what percentage were accurate?", In regression model, the target variable is always continuous. Therefore, if we begin assessing the performance of our model based on various accuracy parameters then the model will be overfitted. The following metrics are used to evaluate the relative performance of different regression models.

1) Performance Metrics

In this study, the key performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Relative Squared Error (RSE), and R-squared (R^2) are chose as the performance metrics (Vandeput, 2023). RMSE is chosen due to its ability to penalize larger errors more significantly, providing a comprehensive measure of overall model accuracy. MAE is included as it offers a straightforward average of absolute errors, providing insights into the typical magnitude of discrepancies between predicted and actual values. MAPE is particularly valuable in our context as it expresses errors as a percentage of the actual values, offering a relative understanding of the prediction accuracy. RSE is utilized for its ability to present model fit in a relative manner, offering insights into the proportion of total variance explained by the model. Lastly, R^2 is added to quantify the proportion of variance in the dependent variable that is predictable from the independent variables, providing a measure of the model's explanatory power. This combination of metrics ensures a holistic evaluation of the proposed model's performance, capturing different facets of error rates, and relative fit, crucial for robust model assessment.

2) Coefficient of Determination

R squared is a statistical measure that indicates how precisely the data fit the regression line which is termed as coefficient of determination (Brown, 2003). R Squared could be termed as ratio of the amount of variation in the dependent variable that could be accounted for by changes in the independent variable, which is computed as follows in equation 1:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(Xi - Yi)^2}{\sum_{i=1}^{m}(\bar{Y} - Yi)^2} \quad (1)$$

Table 3, 4 and 5 depicts the performance metrics of the regression models, on comparing with the other regression models the proposed OptCatB model outperformed with R Square value of 0.54 when PV1Math as target variable, R Square value derived as 0.55 when PV1Science as target variable and with the R Square value of 0.60 when PV1Reading as target variable. Figure 7 connotes the R-Square value of proposed model correlated with selected regression models.

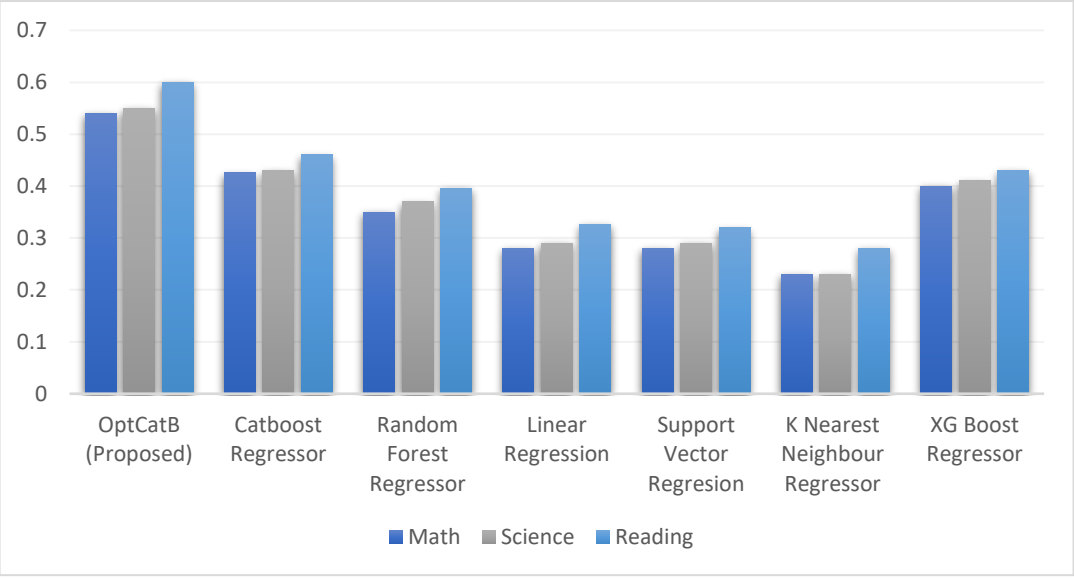

Figure 7: Measuring the R-Square value of OptCatB (Proposed Model)

3) Root Mean Squared Error (RMSE)

The root-mean-squared error (RMSE) is a polynomial ranking method that provides an average measurement of the error (Botchkarev, 2019). Specifically, it is the root square of the sum of all squared deviations between forecast and observation. RMSE computed as follows in equation 2:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((Xi - Yi)^2)} \quad (2)$$

From the performance metrics derived from Table 3, 4 and 5 of the regression models, the proposed OptCatB model performs well with low root mean square error. RMSE of 54.231 when PV1Math as a target variable, RMSE of 54.851 when PV1Science as a target variable and RMSE of 56.040 when PV1Reading as target variable. Figure 8 indicates the range of RMSE score of proposed models correlated with the selected Regression models.

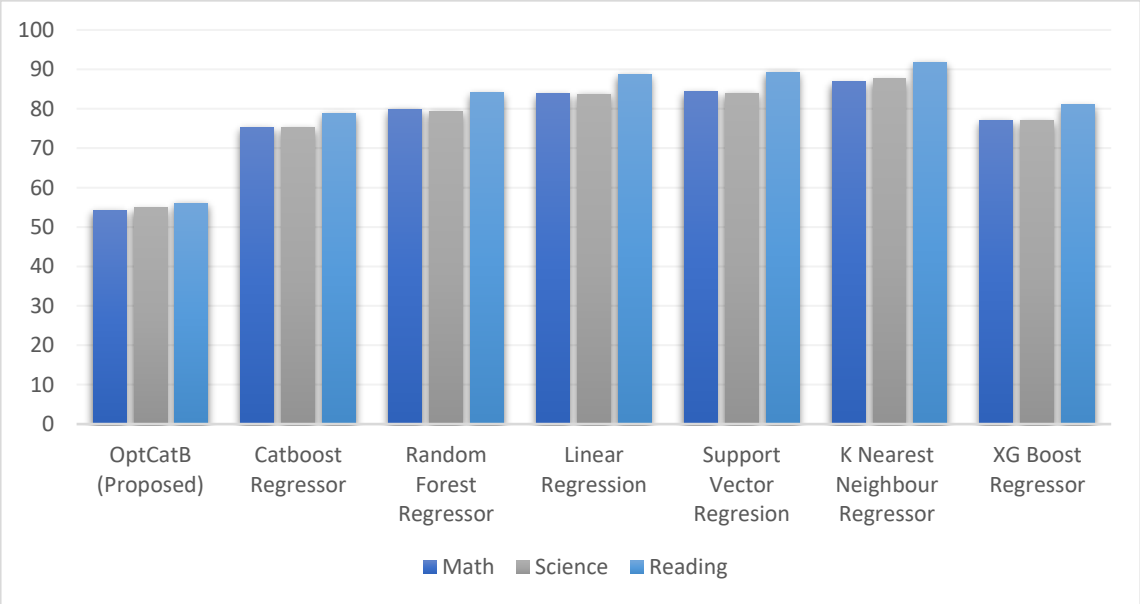

Figure 8: Evaluating the RMSE Score of OptCatB (Proposed Model)

4) Mean Absolute Error (MAE)

Mean Absolute Error is a measurement of the differences in results obtained from comparing paired observations relating to the same occurrence, when outliers reflect tainted elements of the data then the data should be discarded. MAE would provide better results on the normalized dataset, MAE is termed as follows in equation 3:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Xi - Yi| \quad (3)$$

Based on the results from the various regression models shown in Tables 3, 4, and 5 depicts that proposed OptCatB regression model is effective with low MAE value compared with other regression models. The variation in MAE between proposed model and the selected Regression models is seen in Figure 9.

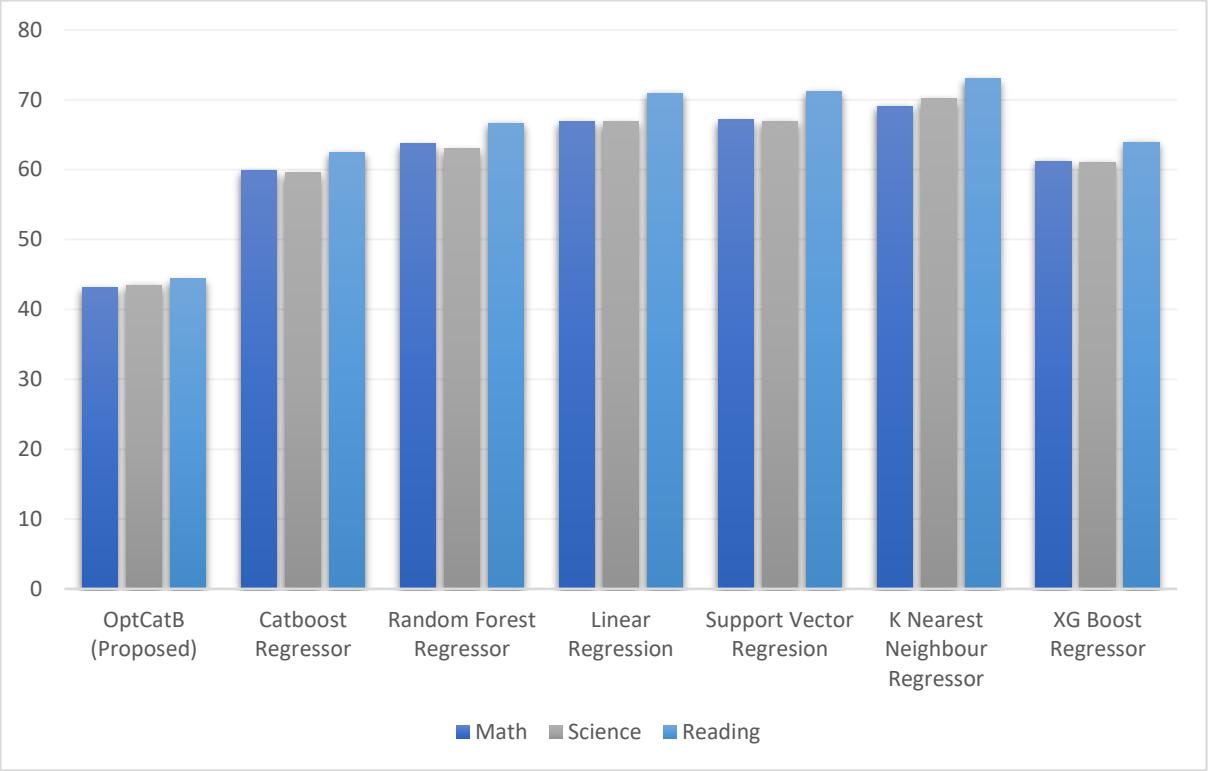

Figure 9: Assessing the MAE value of OptCatB (Proposed Model)

5) Mean Absolute Percentage Error (MAPE)

The MAPE statistic is used to measure the accuracy of a regression model it computes the average absolute percentage error between the observed and projected values. The performance of a regression model could be evaluated by its mean absolute percentage error (Chicco et al., 2021). It is calculated as follows in equation 4:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Yi - Xi}{Yi}\right| \quad (4)$$

Where X_i – absolute value,

Y_i – Predicted value,

n – number of observations

Based on the performance metrics extracted from Tables 3, 4, and 5 of the regression models the effectiveness of the proposed OptCatB regression model is good with low MAPE of 9.931 when PV1Math as target variable, 9.916 when PV1Science as target variable and 10.284 when PV1Reading as target variable. Figure 10 depicts the difference in MAPE among the selected Regression models.

The proposed model Optuna Optimized CatBoost regression outperforms on predicting the academic performance of Immigrant students when stacked up against the selected regression models including Linear Regression, Random Forest Regressor, Support Vector Regression, and k-nearest neighbor Regressor.

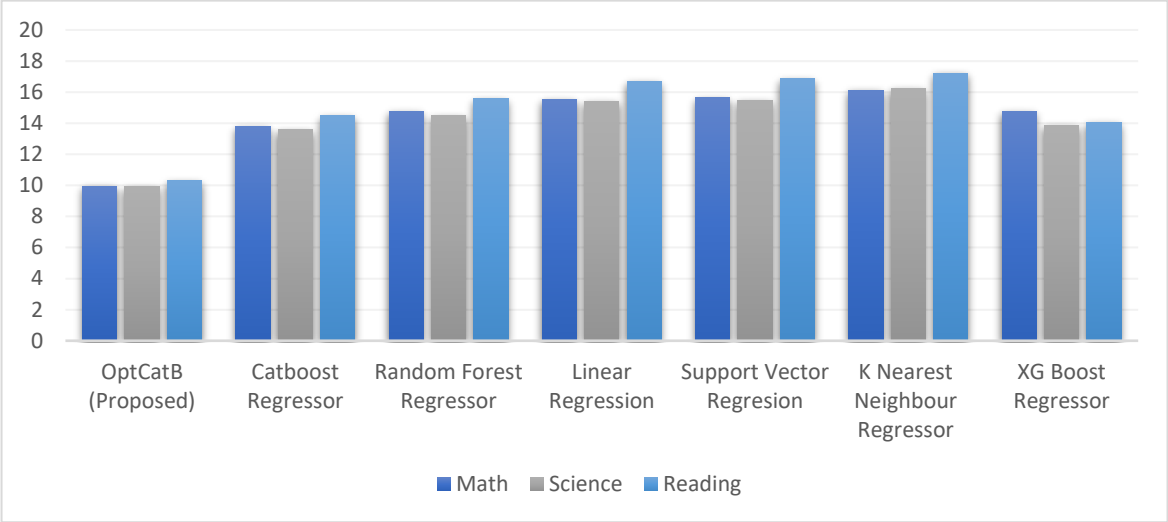

Figure 10: Measuring the MAPE value of OptCatB (Proposed Model)

5 Conclusion and Future Work

The PISA project has sparked a renewed focus on education and the abilities that today's youth must acquire to ensure their own success and the success of their society in the future. Student proficiency in Math, Reading, and Science, as well as the skills necessary to face the challenges of real-world issue are assessed by PISA. Academic success in school may be predicted using machine learning techniques that link various factors like socio-economic status, Parental Education level, educator's perception which in turn helps government policymakers develop strategies to improve quality of education. An exploratory analysis has been carried out to correlate the performance of students and perception of teaching. This research work introduces the concept of a specialized model known as OptCatB. This model predicts the educational performance of immigrant students by using an optimized CatBoost regressor with reading, math, and science as the target variables. After modifying the hyperparameters of the CatBoost algorithm by using the Optuna framework, we trained the model using the optimized parameters. The proposed OptCatB regression model was evaluated with state of art regression models like Random Forest Regressor, Linear Regression, Support Vector Regression and K Nearest Neighbor Regressor. OptCatB regressor outperformed other models with MAPE (9.931, 9.916 and 10.284), RMSE (54.231, 54.851 and 56.04), MAE (43.104, 43.471 and 44.394) and R-Square (0.54, 0.55 and 0.60) and with Maths, Science and Reading as target variable respectively. Future research should be focused on improving prediction proficiency of the model by applying using deep learning techniques.

Author Contributions: Study conception and design: Selvaprabu Jeganathan, Saravanan Parthasarathy; Data curation and visualization: Selvaprabu Jeganathan, A. Abdul Azeez Khan, K. Javubar Sathick; Analysis and interpretation of results: Selvaprabu Jeganathan; Draft manuscript preparation: Selvaprabu Jeganathan; Supervision, Validation, Reviewing: Arun Raj Lakshminarayanan.

Availability of Data and Material: The source dataset analysed during the current study is available at [https://www.oecd.org/pisa/data/2018database/]. The customized dataset generated and/or analysed during the current study is available from the corresponding author on reasonable request.

Funding: This research received no external funding.

Conflict of Interest: The authors declare no conflict of interest.

Acknowledgments: We would like to thank the 'Organisation for Economic Co-operation and Development' for providing publicly available data which was used in this study.

References

[1] Akther, A., & Robinson, J. (2014). Immigrant Students' Academic Performance in Australia, New Zealand, Canada and Singapore. *Australian Association for Research in Education, Joint AARE-NZARE 2014 Conference, Brisbane*, 1-7.

[2] Andon, A., Thompson, C.G., & Becker, B.J. (2014). A quantitative synthesis of the immigrant achievement gap across OECD countries. *Large-scale Assessments in Education*, *2*(1), 1-20.

[3] Beaulac, C., & Rosenthal, J.S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, *60*(7), 1048-1064.

[4] Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk-predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, *11*(3), 1–41.

[5] Bharathi, C., & Rekha, D. (2023). Load Forecasting for Demand Side Management in Smart Grid using Non-Linear Machine Learning Technique. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA)*, *14*(1), 200-214.

[6] Botchkarev, A. (2018). Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio. http://dx.doi.org/10.2139/ssrn.3177507

[7] Brown, J.D. (2003). The coefficient of Early Detection of Students at Risk-predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods determination. *JALT Testing & Evaluation SIG Newsletter*, *7*(1), 14-16.

[8] Browning, M.H., & Rigolon, A. (2019). School green space and its impact on academic performance: A systematic literature review. *International journal of environmental research and public health*, *16*(3), 429. https://doi.org/10.3390/ijerph16030429

[9] Cameron, S. (2012). Education, urban poverty and migration: Evidence from Bangladesh and Vietnam. https://doi.org/10.18356/b21a829f-en

[10] Camgozlu, Y., & Kutlu, Y. (2023). Leaf Image Classification Based on Pre-trained Convolutional Neural Network Models. *Natural and Engineering Sciences*, *8*(3), 214-232.

[11] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623.

[12] Coleman, C., Baker, R.S., & Stephenson, S. (2019). A Better Cold-Start for Early Prediction of Student At-Risk Status in New School Districts. *International Educational Data Mining Society,* 732-737.

[13] Cortes Pascual, A., Moyano Muñoz, N., & Quílez Robres, A. (2019). The relationship between executive functions and academic performance in primary education: Review and meta-analysis. *Frontiers in psychology*, *10*, 449759. https://doi.org/10.3389/fpsyg.2019.01582

[14] Costa-Mendes, R., Cruz-Jesus, F., Oliveira T., & Castelli, M. (2021). Machine learning bias in predicting high school grades. *Emerging Science Journal*, *5*(5), 576-597.

[15] Danh, N.T. (2020). A compact dual polarized ring slot loaded patch antenna for Navic applications. *National Journal of Antennas and Propagation (NJAP)*, *2*(1), 1-6.

[16] Feldman-Maggor, Y., Barhoom, S., Blonder R., & Tuvi-Arad, I. (2021). Behind the scenes of educational data mining. *Education and Information Technologies*, *26*(2), 1455-1470.

[17] Hooshyar, D., Margus, P., & Yeongwook, Y. (2020). Mining Educational Data to Predict Students' Performance through Procrastination Behavior. *Entropy*, *22*(1), 12. https://doi.org/10.3390/e22010012

[18] Hussain, S., & Khan, M.Q. (2023). Student-per formulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of data science*, *10*(3), 637-655.
[19] Immigration by Country 2024. (n.d.). https://worldpopulationreview.com/country-rankings/immigration-by-country.
[20] Inoue, K., Seeman, T.E., Nianogo, R., & Okubo, Y. (2023). The effect of poverty on the relationship between household education levels and obesity in US children and adolescents: An observational study. *The Lancet Regional Health–Americas*, *25*. https://doi.org/10.1016/j.lana.2023.100565
[21] Jonnerby, J., Brezger, A., & Wang, H. (2023). Machine learning based novel architecture implementation for image processing mechanism. *International Journal of Communication and Computer Technologies (IJCCTS)*, *11*(1), 1-9.
[22] Juma, J., Mdodo, R.M., & Gichoya, D. (2023). Multiplier Design using Machine Learning Algorithms for Energy Efficiency. *Journal of VLSI Circuits and Systems*, *5*(1), 28-34.
[23] Khan, A., & Ghosh, S.K. (2021). Student performance analysis and prediction in classroom learning: a review of educational data mining studies. *Education and information technologies*, *26*(1), 205-240.
[24] Levy, J., Mussack, D., & Fischbach, A. (2020). Contrasting classical and machine learning approaches in the estimation of value-added scores in large-scale educational data. *Frontiers in psychology*, *11*, 561534. https://doi.org/10.3389/fpsyg.2020.02190
[25] Lewin, A.C. (2023). Poverty, affluence and homeownership among working-age immigrants in Israel. *Population, Space and Place*, *29*(1), e2601. https://doi.org/10.1002/psp.2601.
[26] Libretexts. (2021). 13.4: Education as signalling. Social Sci LibreTexts. https://socialsci.libretexts.org/Bookshelves/Economics/Principles_of_Microeconomics_(Curtis_and_Irvine)/05%3A_The_Factors_of_Production/13%3A_Human_capital_and_the_income_distribution/13.04%3A_Education_as_signalling
[27] Liloja & Ranjana, P. (2023). An Intrusion Detection System Using a Machine Learning Approach in IOT-based Smart Cities. *Journal of Internet Services and Information Security (JISIS)*, *13*(1), 11-21.
[28] Livieris, I.E., Drakopoulou, K., Tampakas, V.T., Mikropoulos, T.A., & Pintelas, P. (2019). Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of educational computing research*, *57*(2), 448-470.
[29] Ngepah, N., Makgalemele, T., & Saba, C.S. (2023). The relationship between education and vulnerability to poverty in South Africa. *Economic Change and Restructuring*, *56*(1), 633-656.
[30] Poskitt, J. (2018). Immigrant student achievement and education policy in New Zealand. *Immigrant student achievement and education policy: Cross-cultural approaches*, 175-193.
[31] Prenkaj, B., Velardi, P., Stilo, G., Distante D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. *ACM Computing Surveys (CSUR)*, *53*(3), 1-34.
[32] Ramaswami, G., Susnjak, T., & Mathrani, A. (2022). On developing generic models for predicting student outcomes in educational data mining. *Big Data and Cognitive Computing*, *6*(1), 6. https://doi.org/10.3390/bdcc6010006
[33] Rastrollo-Guerrero, J.L., Gómez-Pulido, J.A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied sciences*, *10*(3), 1042. https://doi.org/10.3390/app10031042
[34] Rebai, S., Yahia, F.B., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences*, *70*, 100724. https://doi.org/10.1016/j.seps.2019.06.009
[35] Ross, S. (2023). What is the human capital theory and how is it used? Investopedia. https://www.investopedia.com/ask/answers/032715/what-human-capital-and-how-it-used.asp

[36] Srinivas, P., & Katarya, R. (2022). hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. *Biomedical Signal Processing and Control*, *73*, 103456. https://doi.org/10.1016/j.bspc.2021.103456.
[37] Srinivasa Rao, M., Praveen Kumar, S., & Srinivasa Rao, K. (2023). Classification of Medical Plants Based on Hybridization of Machine Learning Algorithms. *Indian Journal of Information Sources and Services (IJISS)*, *13*(2), 14–21.
[38] Team, C. (2023). Social mobility. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/economics/social-mobility/
[39] Vandeput, N. (2023). Forecast KPI: RMSE, MAE, MAPE & BiAS | Towards Data Science. Medium. https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d
[40] Welch, A. (2018). Immigrant student achievement and education policy in Australia. *Immigrant Student Achievement and Education Policy: Cross-Cultural Approaches*, 155-173.
[41] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z.
[42] Young, N.T., & Caballero, M.D. (2021). Predictive and explanatory models might miss informative features in educational data. *arXiv preprint arXiv:2103.14513*.
[43] Zhou, M., Wang, L., Wu, H., Li, Q., Li, M., Zhang, Z., & Zou, Z. (2022). Machine learning modeling and prediction of peanut protein content based on spectral images and stoichiometry. *LWT*, *169*, 114015. https://doi.org/10.1016/j.lwt.2022.114015.

Authors Biography

Selvaprabu Jeganathan Obtained his bachelor's degree in information technology from the SASTRA University, Thanjavur in 2005. He completed his master's degree in computer science and engineering from Anna University, Chennai in 2015. He obtained his Postgraduate degrees in Geography and E-Business Management from Tamil University, Thanjavur and Annamalai University. He is currently pursuing his doctorate from B.S. Abdur Rahman Institute of Science and Technology in the field of Educational Data Mining. He has 15 years of experience in the software industry. At present, he is affiliated to B.S. Abdur Rahman Crescent Institute of Science and Technology, as a part-time research scholar. He has published 8 research articles in international conferences and journals.

Arun Raj Lakshminarayanan Embarked on his academic journey with a bachelor's degree in computer science and engineering from Anna University in 2006. He continued his pursuit of knowledge by completing his master's degree in computer science and engineering, also from Anna University, in 2010. His dedication to academia and research culminated in the successful completion of a Ph.D. in Information and Communication Engineering from Anna University in 2017. He currently serves as an "Associate Professor" at B.S. Abdur Rahman Crescent Institute of Science and Technology. The breadth of his knowledge and the impact he has had in academia highlight his dedication to computer science and engineering. His scholarly significance is further enhanced by the publication of 40 research articles in international conferences and journals.

Saravanan Parthasarathy Obtained his bachelor's degree in computer science and engineering from the Madurai Kamaraj University, Madurai in 2003. He completed his master's degree in computer science and engineering from Anna University, Chennai in 2015. He obtained his Postgraduate degrees in Psychology, Medical Sociology and Business Management from TNOU, Madras University, Alagappa University and Annamalai University. He is currently pursuing his doctorate from B.S. Abdur Rahman Institute of Science and Technology in the field of Crime Informatics. He has 13 years of experience in the software industry. At present, he is affiliated to B.S. Abdur Rahman Crescent Institute of Science and Technology, as a full-time research scholar. He has published 15 research articles in international conferences and journals.

A. Abdul Azeez Khan Embarked on his academic journey by earning a bachelor's degree in computer science from the University of Madras in 2002. Building upon his foundation, he pursued a Postgraduate degree in Computer Science from the same institution in 2004. In 2008, he achieved his M.Phil. degree from Periyar University. Driven by a thirst for knowledge, he went on to complete his doctorate from B.S. Abdur Rahman Institute of Science and Technology in 2018. With a cumulative experience of 13 years in both teaching and industry, Dr.A. Abdul Azeez Khan currently serves as an Associate Professor in the Department of Computer Applications at B.S. Abdur Rahman Crescent Institute of Science and Technology. His academic prowess is proved through his publication of 7 research articles in international journals and the presentation of 6 papers in international conferences.

K. Javubar Sathick Received his Doctorate in Computer Science from B.S. Abdur Rahman Institute of Science and Technology in 2018. Prior to this, he completed his Postgraduate degree in Computer Applications from Anna University in 2008. With a robust background spanning 13 years in both teaching and industry, Dr.K. Javubar Sathick currently holds the position of Associate Professor in the Department of Computer Applications at B.S. Abdur Rahman Crescent Institute of Science and Technology. He has published 10 research articles in international journals. His research interests encompass Knowledge Management, E-Learning, Web Mining, Web Applications, Big Data, and Information Retrieval.