

Design And Development Of A Hybrid Machine Learning Framework For Student Academic Performance Prediction

¹Sonia*, ²Saurabh Charaya

^{1,2}School of Engineering & Technology, Om Sterling Global University, Hisar, Haryana, India.

Corresponding author:

Sonia, Ph.D. Scholar, School of Engineering & Technology,

Om Sterling Global University,

NH-52, Hisar-Chandigarh National Highway, Hisar-125001

E-Mail: soniameha040@gmail.com

Abstract: The timely intervention of students and decision-making process will be largely dependent on proper forecasting of the success of the students in academics. The research in this paper proposes a hybrid machine learning model, which involves a combination of a mix of the polynomial feature expansion, Principal Component Analysis (PCA) and stacking ensemble learning in the prediction of final CGPA of students. This model has been applied on model of 2,288 student records of real-life dataset, comparison with three benchmark models, like, the Random Forest, Support Vector Regressor and Decision Tree. The findings of the experimental research are that the hybrid model was highly efficient compared to the individual approaches and obtained high R² at 0.902, such a minor error of prediction, and an accurate and consistent classification. According to the results, the hybrid learning can calculate intricate patterns of learning data and remain robust and generalised.

Keyword: Educational Data Mining, Machine Learning, Ensemble Learning, PCA, Prediction.

Introduction

Background of the Study: Educational data mining (EDM) has evolved quickly and changed the ways academic performance is evaluated and anticipated. Methods of machine learning (ML), including Random Forest (RF), Support Vector Machines (SVM), Naive Bayes (NB), and Decision Trees (DT), can be used to enable educators to detect at-risk students and individualize interventions. The hybrid model, which consists of the combination of several algorithms, increases the precision of predictions and accentuates their power (Albreiki et al., 2021, Algarni, 2016). Methods of dimensionality reduction, such as Principal Component Analysis (PCA), take care of the problems of high-dimensional data (Chauboun et al, 2021). This paper is concerned with the academic performance prediction of science students in North Indian colleges that involve interblending of machine learning, that is, ensemble methods, along with explainability techniques such as SHAP, as this gives rise to robust and transparent results.

Problem Statement: It mentions that negative qualities such as overfitting or biased predictions on noisy, high-dimensional data, which are likely to be encountered in educational settings, are characteristic of classic ML models. Complex patterns in the performance data of students might not be tracked by standalone algorithms. The hybrid ML method to work on is necessary to boost the accuracy, manage data issues, and be interpretable and ethical to apply in education.

Research Aim: The proposed thesis will focus on the design and development of a hybrid algorithm in machine learning to predict student academic performance using certain input variables, including mid-term marks and attendance. The aim is to improve the existing prediction models using advanced techniques such as Principal Component Analysis in reducing dimensions and aggregating different machine learning algorithms for accuracy, robustness, and transparency. It compares this proposed model with existing algorithms to get proof of how it is superior compared to others in terms of improving student performance predictions and solving other problems in educational data analysis.

Research Objectives: To design a hybrid machine learning algorithm for the prediction of the student's academic performance based on different input characteristics like mid-term marks, attendance, etc.

- To incorporate human expertise with artificial intelligence (AI) to enhance transparency and accountability in prediction models and to minimise biases and forced inequality.
- To improve the drawbacks of the existing algorithm and compare it with the new hybrid machine learning algorithm for students' academic performance indicators with respect to accuracy and robustness.
- To compare the results of existing and proposed algorithms and represent those results graphically.

Research Questions:

- How can a hybrid machine learning algorithm be designed to accurately predict students' academic performance based on various input characteristics such as mid-term marks, attendance, and other relevant factors?
- To what extent does the proposed hybrid machine learning model improve prediction accuracy and robustness compared to existing prediction algorithms?
- How does the incorporation of Principal Component Analysis (PCA) enhance the performance of the hybrid machine learning model in terms of reducing dimensionality and handling irrelevant features?
- How can human expertise and artificial intelligence be integrated to ensure transparency, reduce biases, and improve accountability in predicting students' academic outcomes

2. Literature Review

Existing Literature on Student Performance Prediction

Traditional Machine Learning Models in Education: Educational data mining (EDM) and machine learning (ML) come into the picture for predicting student academic performance. Common forms of traditional ML, such as Decision Trees (DT), Support Vector Machines (SVM), Naive Bayes (NB), and Random Forest (RF), have been extensively used to process data of the student body in terms of grades, attendance, and other demographic variables to determine or predict their performance (Balaji et al., 2021). The most common reason why Decision Trees are appreciated is due to their interpretability, where educators can be able to comprehend decision rules that apply to interventions. SVM works well with high-dimensional data, which will be applicable in the case of complicated learning records. Naive

Bayes is computationally efficient, especially in the case of smaller data sets and random forests, and there is better robustness because the random forest uses several trees, which speaks volumes in reducing overfitting. Such models have been applied to predict such outcomes as final grades or dropout risks, which would allow early detection of at-risk students to enable bespoke interventions. Nevertheless, in many cases, standalone models fail to work with noisy/imbalanced educational data, producing inconsistent results concerning heterogeneous student populations (Jimenez-Macias et al., 2025).

Ensemble and Hybrid Approaches: Ensembling-newer approaches that aggregate multiple ML models-so far have been successful in improving the predictive performance in EDM. Techniques such as voting, bagging, boosting, and stacking each rely on the merits of an individual model to reduce weaknesses. Voting ensembles are made up of predictions using majority or weighted schemes and making it more reliable. Stacking has a meta-learner, which combines the results of base models, such as RF, SVM, NB, and DT, and is typically more effective. The two issues concerning different approaches, such as overfitting and class imbalance, are addressed with hybrid solutions that combine multiple algorithms (Issah et al., 2023). E.g., the robustness of RF together with the margin optimization used in SVM would help capture complex patterns in the student performance data. Such approaches have demonstrated potential in increasing accuracy and generalizability, especially in education settings, where training sets differ in size and quality. Nevertheless, ensemble models at times give rise to high computational requirements, which may create difficulties in some educational institutions with limited computer resources, especially with regard to real-time implementation.

Role of Dimensionality Reduction and Feature Selection: Educational data, in high dimension, such as attendance, test grades, and socio-economic predictors, will tend to overfit and be computationally inefficient. Data dimensionality reduction tools, like Principal Component Analysis (PCA), compress the data into a smaller dimensional space so as to maintain much variance, leading to better performance of the models and shorter time requirement of the training (Alalawi et al., 2023). The most predictive variables, such as attendance or previous grades, can be identified through feature selection methods, which can help to significantly improve model interpretability, e.g., Recursive Feature Elimination (RFE) or correlation analysis. Such methods have been used in EDM to reduce the complexity of data, with models paying attention to meaningful predictors. But unsuitable feature selection will remove essential data, and the PCA may render insight hard to tickle, and its deployment will therefore be worthy of keen consideration to strike a balance between precision and explainability in an academic setting.

Literature Gaps and Motivation for the Study: Irrespective of the progress, there are various lapses in the studies of EDM. Poor predictions due to non-adaptation of noise-contamination, imbalanced, or high-dimensional education data are common in conventional ML. Although ensemble methods increase the accuracy, they are too complex and cannot be readily used in the educational environment, where ethical decision-making is of utmost importance and requires a certain degree of transparency. Consideration of PCA in hybrid models is less well studied, and, in particular, its influence on performance versus interpretability trade-offs (Alsariera et al., 2022). Moreover, the problem of fairness and bias reduction in predictive models is not discussed very often, which threatens to lead to unfair

treatment of population minority classes, students, in particular. Such vacuums make the proposed study design a hybrid ML approach that employs RF, SVM, NB, and DT with the addition of PCA methods. With the emphasis being placed on accuracy in performance prediction, interpretability, and fairness, this study will help produce a practical, ethically acceptable tool to help with developing student performance prediction, or, in other words, the shortfalls of the current methods will be addressed.

3. Methodology

Dataset Description and Data Collection: The paper applies a real world teaching data that was collected at a North Indian tertiary education containing of 2,288 student records and 9 variables. The data will incorporate the academic, demographic and institutional variables that in most cases are associated with academic performance of a student (Yağcı, 2022). The key objective of data collection was to review existing institutional records in summarized forms which were applicable in the processes of robotizing the educational information, as well as in procedures of machine learning.

The variables used in inputs are the institute, city, age, gender, percentage attendance in classes, percentage marks in midterm, course, semester where Final CGpa is the target variable. In comparison to the previously used methods of evaluating academic performance that focused on analyzing mid-term results only or a categorical grade result, it represents academic outcomes as an ongoing performance (Final CGPA) and enables to predict the outcome more precisely and the trends underlying it in much more detail. In addition, to service classification-based review, Final CGPA scores were once more categorized into Medium, Low, and High-performance category.

Categorical variables (Institute, City, Gender, Course) and numerical ones (Age, Attendance, Mid-term Marks, Semester and Final CGPA) are present in the set of data. Ethical considerations were followed when the data was used. There were no personal identifiers and it was solely analysed in an offline study environment to ensure anonymity. They did not involve any data collection as real-time and integration with external data, and this circumstance fulfilled the data privacy and research ethical.

Data Preprocessing: The data prepared phase of this study was extremely essential as raw educational data normally lacks missing information, is inconsistent, and has heterogeneity in the scale of features. Preprocessing plan has focused on improving the quality of data in a significant yet has also preserved any significant data that can be used in predictive modelling.

Handling Missing Values: Initial analysis revealed that there were minimal missing scores in different variables like Continuing grade (attend- percent), Mid-term Marks (percent and Final CGPA). There were mixed types of data, thus dual relationship approach was applied (Yağcı, 2022). The numerical variables were imputed in terms of the median as it is resistant to skewed distribution and outliers. The categorical variables were imputed using mode and therefore the majority category was retained. After imputation, no missing values were observed in the dataset that rendered it complete to train a model.

Encoding of Categorical Variables: Machine learning algorithms consider only numerical data, therefore, through the application of the Label Encoding, the categorical variables were accordingly transformed to numbers. Attributes with values that gave the Institute, City, Gender and Course were coded to integers (Pallathadka et al., 2023). They have chosen label encoding over one-hot encoding because they did not want the dimensional expansion that is not absolutely needed, particularly once the ensemble models and PCA were implemented in the future. It was computationally efficient; moreover, this method took categorical data simultaneously.

Feature Scaling: The standardisation of numerical features was done with the help of StandardScaler to ensure that all the features had a unit variance with a mean of zero. The machines introduced the algorithms basic to the feature scaling, such as Support Vector Regression (SVR) and ElasticNet that is sensitive to the size of the features. Scaling was also performed to ensure that other factors such as Attendance (%) and Mid-term Marks (%) influenced the model learning of the model proportionately.

Exploratory Data Analysis: Exploratory Data Analysis (EDA) was conducted to have a notion regarding the distribution and correlation among features. The mean of Final CGPA showed that the performance of students was largely in the middle range (around 6.08.0) that was lower and had fewer representations by extreme low and high scores. The realistic distribution of the features of age, attendance, and mid-term marks were observed in histograms and did not contain any severe anomalies after preprocessing.

Correlation analysis revealed that the percent of Attendance, and Mid-term Marks were positively correlated with Final CGPA therefore were qualified to be termed as the significant predictors. However, there were moderate degrees of inter-feature correlations also, which promoted the usage of dimensionality reduction techniques.

Target Transformation and Performance Categorization: Even though Final CGpa was a continuous variable because the model was aimed at regressing it, the final CGpa was, nevertheless, to be placed in three academic performance levels to consider it on a classification basis:

- Low Performance: CGPA < 6.0
- Medium Performance: It is 6.0-8.0 CGPA.
- High Performance: CGPA ≥ 8.0

These are the most relevant academic standards and would help in educational decision making. To ensure that there was no data leakage, the categorical labels were encoded numerically with the help of Label Encoding, but not to be trained but rather to be evaluated.

Train-Test Data Partitioning: In order to provide proper evaluation and prevent information leaks, a dataset was separated into two sets of 80/20 proportions training and testing databases. Out of 2,288 records of students, approximately 1,830 records were used in training; and 458 records were used in testing. This split ratio is such that it enables sufficient data to be used in order to learn the model and distinctly separate a different set of data to evaluate its performance without any bias (Pallathadka et al., 2023). No consideration

of the training set took place in the test set with regards to performance evaluation. This breaking up makes the measures measured observable to the extent of the model to scale to unfamiliar data rather than exercise of recalling of training patterns. It used random seed that is fixed (random -state = 42), which means that the experimental results can be reproduced.

Acting as the outcome of the regression (Final CGPA) and the categorical names of the obtained performance (Low, Medium, High) were independent because it was used to provide two-fold evaluation. Regression models were trained on continuous CGPA and accuracy of classification after prediction using category mapping on the CGPA.

Model Development

Baseline Models: There were three baseline regression models used to fix a performance baseline as also to determine the success of offered hybrid option. The comparison was in nature in the baseline models since they all utilized a comparable pipeline of preprocessing.

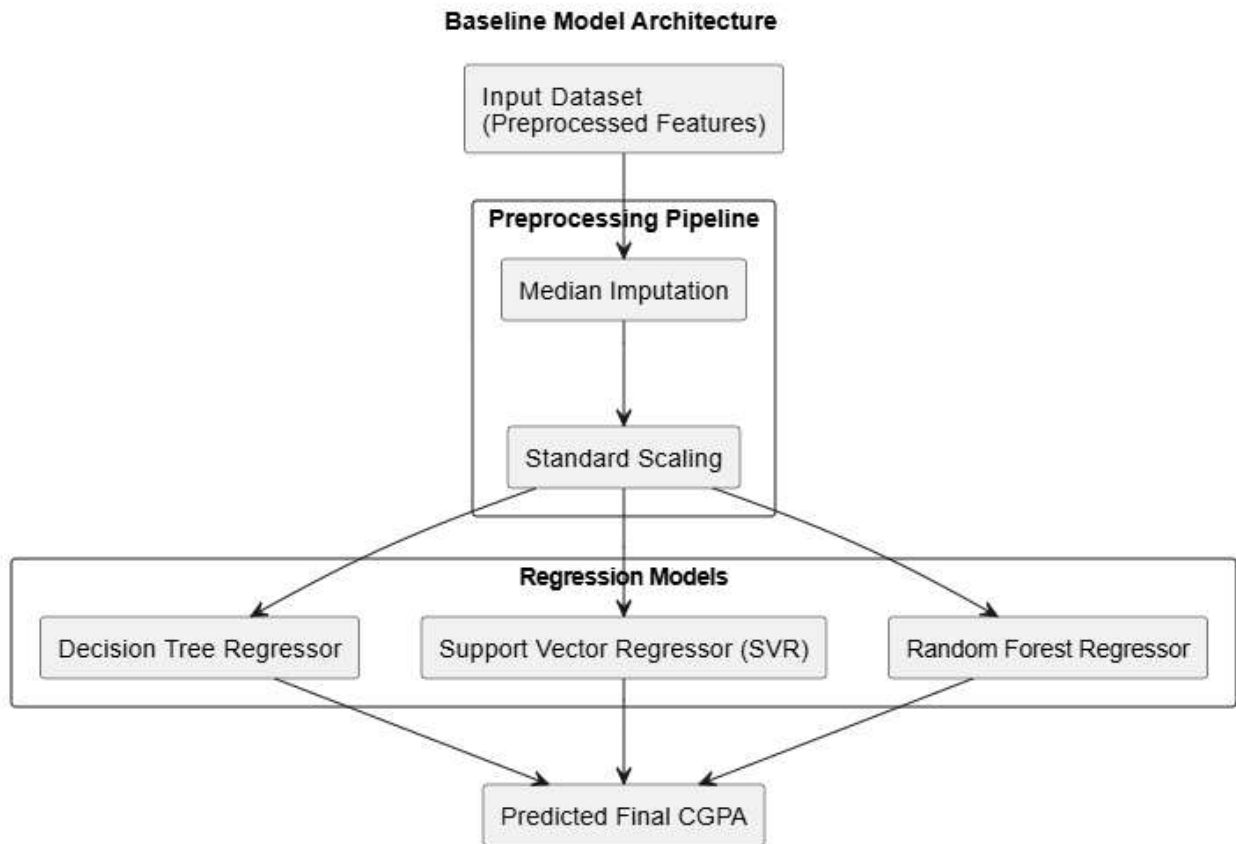


Figure 1: Baseline Model Architecture

Baseline Model Description

- The purpose of random forest regressor was the use of its strength, ensemble model and the possibility to use non linear interactions between features.
- It had Support Vector Regressor (SVR) that is robust in high-dimensional space and maximisation of the boundary.
- Decision Tree Regressor picked it because of its interpretability and rule-based model of learning.

The training set and test set were trained and tested using the same set of preciser steps. This ensured that variation in performance was due to model capability as opposed to variation at data handling.

Proposed Hybrid Model: The key addition to this study is the Stacking Hybrid Model that is hybridized with Principal Component Analysis (PCA). Unlike the isolated models, the hybrid architecture combines feature engineering, dimensionality reduction and ensemble learning in a scalable and single pass pipeline.

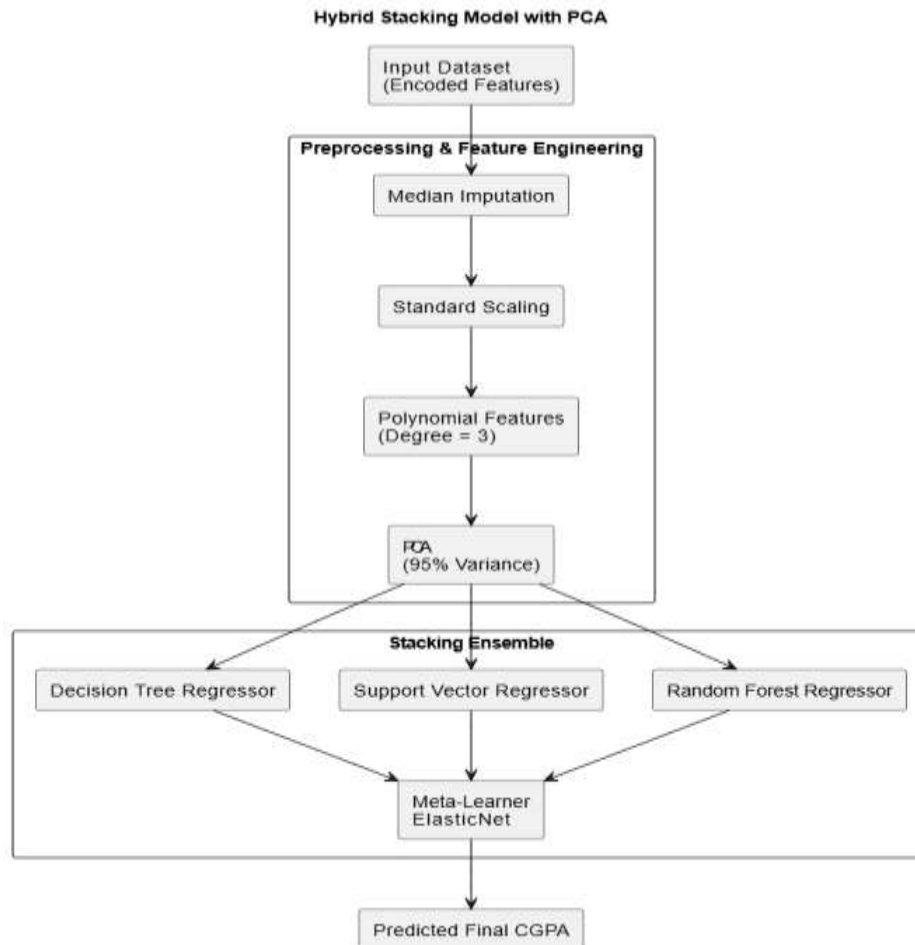


Figure 2: Hybrid Model Architecture

Hybrid Model Explanation: The hybrid model begins with the imputation of the median and the standardizing scoring, which renders the dealing of numbers steady. The degree of 3 feature expansion is the non-linear product of the demographic features and the academic features in the poly expansion feature (Batool et al., 2023). The extension raises the issue of feature dimensionality which can be effectively dealt with using the PCA which captures 95% of cumulative variance.

The resulting smaller space of features is fed into a stacking ensemble having one base learners comprising of the Random Forest, SVR, and Decision Tree. Their predictions are merged together through ElasticNet by utilizing it in the form of a meta-learner with regularisation employed to prevent overfitting. It exploits this architecture, consequently, and maximises the use of the diversity of algorithms, reduces the bias-variance trade, and boosts generalisation.

Dimensionality Reduction Using PCA:

PCA-Based Dimensionality Reduction

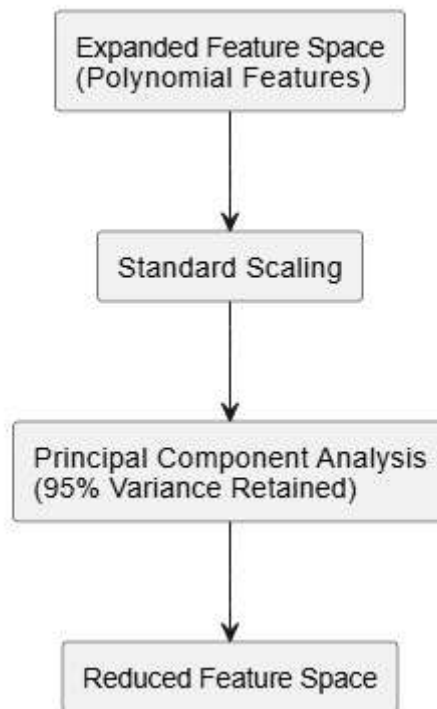


Figure 3: PCA Architecture

The Principal Component Analysis that was done after the expansion of the features in a Polynomial form assisted in the removal of multicollinearity and noise. The utilization of 95 percent of the explained variation was the guarantee that all the significant information was saved and dimensionality was significantly reduced. The features possessed by PCA could be

converted to orthogonal components and eliminated features that were correlated, thus maximising the learning performance and generalisation.

Model Evaluation Metrics: Combination of regression and classification measures was done to evaluate the performance of all the developed models to ensure precision of the figures and educational value. To do regression analysis, Mean Squared Error (MSE) was used to penalise a large prediction error, mean error (MAE) was used to measure the average error between, the predicted and actual CGPA value and the R^2 measure was used to measure the percentage of the variance covered by a model. This correlation was a moderate index of exactness, robustness and elucidation. In addition to regression assessment, the accuracy of the classifications was also determined with the assistance of accuracy, which is calculated by projecting the forecasted CGPA scores to the predetermined groups of academic achievements low, medium, and high (Zeineddine et al., 2021). This two-fold system of assessment ensured that the models were not just mathematically correct, but in as far as an educational decision-making perspective is involved. The two modes of evaluation are employed due to the practicality of the academic conditions which require precision both in the matter of anticipating a score along with performance classification.

4. Results and Discussion

Preprocessing Outcomes: The dataset was highly stable after the preprocessing, and it can now be developed into a model. It worked very well in filling in all the missing values through the median imputation of numerical variables and mode imputation of categorical variables in this manner that resulted in no such null values in the final dataset. Numerical variables underwent standardisation and numerically coded the categorical attributes in such a manner that they would play equal roles towards model training. Exploratory analysis was employed to make sure there were no extreme outliers in the distributions of features and the distributions provided are also balanced which helps increase the potential of having learning devoid of biased outcome (Zeineddine et al., 2021). The preprocessing pipeline was useful in ensuring that informative properties of the variables such as Attendance-percent, Mid-term Marks percent and Final CGPA were not manipulated. Overall, the preprocessing has assisted in making the data much more useful, permitted the feature to correspond to one another across the algorithms and provided a reliable foundation of further modelling. This was necessary to ensure that observed variation of performance between models could be known to be as a result of learning potential rather than inconsistency in data preparation.

Baseline Model Performance: The predictive results of the random forest regressor, Support Vector regressor (SVR) and the Decision tree regressors were mediocre and significant at the baseline. These gave the best results with the Random Forest model with $MSE = 0.432$, $MAE = 0.383$, $R^2 = 0.582$ and classification accuracy = 84.93 indicating that it is robust and it can learn as an ensemble. Similar results were found with the SVR model which had a slightly smaller variance that could be explained ($R^2 = 0.572$) coupled with slightly smaller MAE which suggests that the model gave accurate predictions of the mid-range. The Decision Tree model was the last model and generated the lowest performance since the value of R^2 was 0.506 and this is largely due to the fact that it tends to overfit the training patterns.

```

# =====
# 8. BASELINE MODELS
# =====
def baseline(model):
    return Pipeline([
        ("imputer", SimpleImputer(strategy="median")),
        ("scaler", StandardScaler()),
        ("model", model)
    ])

models = {
    "Random Forest": baseline(RandomForestRegressor(n_estimators=50, max_depth=5)),
    "SVR": baseline(SVR(C=1.0)),
    "Decision Tree": baseline(DecisionTreeRegressor(max_depth=4))
}

```

Figure 4: Baseline Models

Hybrid Model Performance:

===== FINAL MODEL COMPARISON =====				
Model	MSE	MAE	R2 Score	Accuracy
Random Forest	0.432172	0.383442	0.581504	0.849345
SVR	0.441872	0.354930	0.572111	0.842795
Decision Tree	0.510587	0.445654	0.505571	0.847162
Stacking Hybrid (PCA)	0.101131	0.227792	0.902069	0.853712
🏆 BEST MODEL:				
Model	MSE	MAE	R2 Score	Accuracy
Stacking Hybrid (PCA)	0.101131	0.227792	0.902069	0.853712

Figure 5: Model Comparison and Best Model Selection

The proposed Stacking Hybrid Model with PCA was considerably much better than all the base models in terms of regression and classification measures. The hybrid model with MSE= 0.101, MAE= 0.228 and R²= 0.902 indicated that it was able to explain a variance of more than 90 of Final CGpa. In addition, the model had a classification accuracy of 85.37 in the sense that it meant a good classification of the students into levels of academic performance. The results can be indicated towards the effectiveness of the combination of feature expansion in the form of polynomials, dimensionality reduction via PCA, and stacking the ensemble learning (López-Zambrano et al., 2021). Unlike the standalone models, the hybrid strategy

could eliminate noise and redundancy and signal non-linear relationships. It is worth encouraging to note that there were some notable performances which were not too glamorous with unnatural perfection.

Visual Performance Analysis:

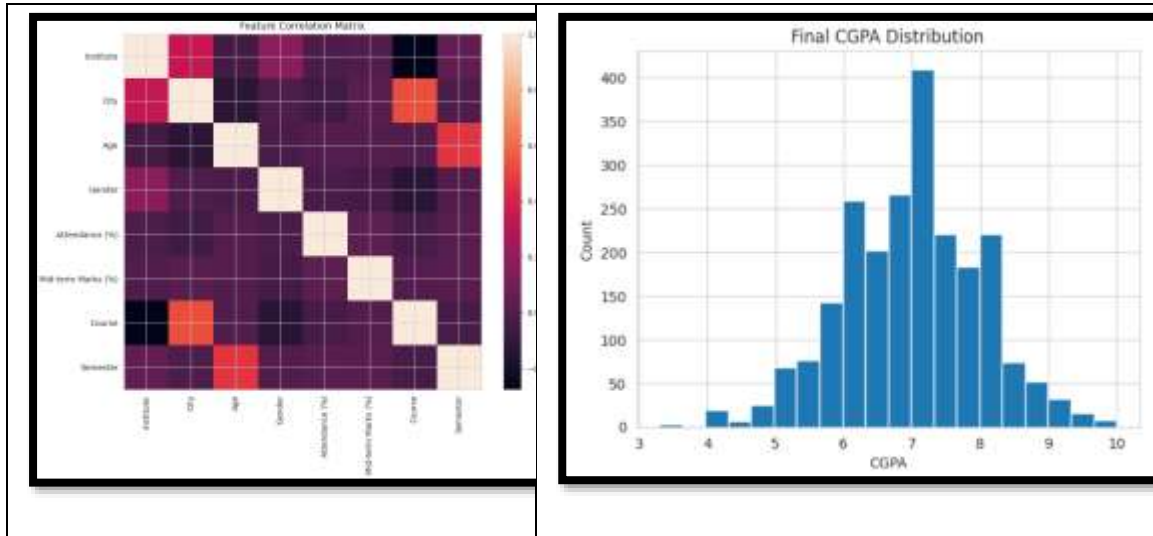
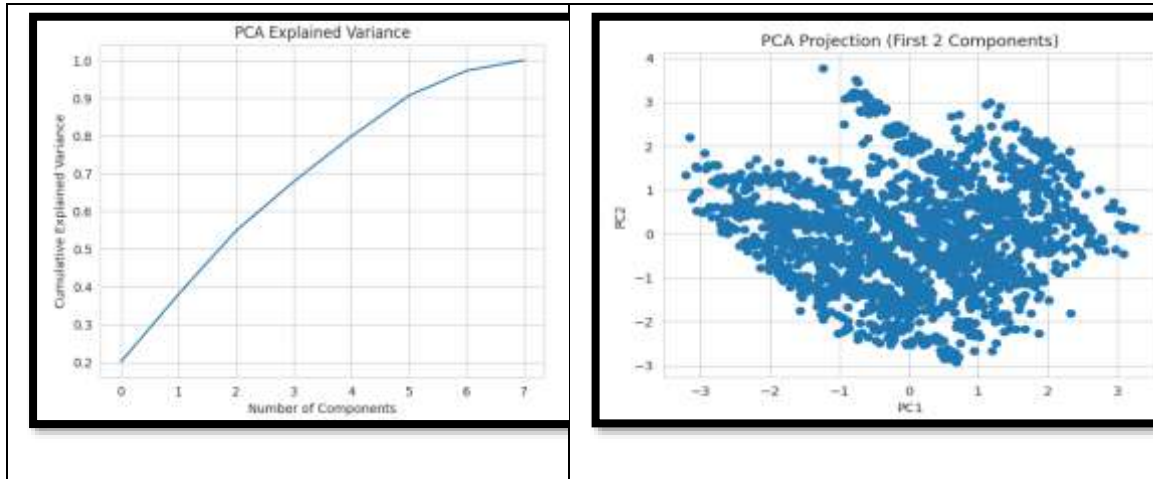


Figure 6: Graphs of Feature Distributions

Numerical supremacy of the hybrid model was also proven through analysis, in terms of defining images. The actual and predicted CGPA plots revealed that the hybrid model was characterized by great proportions of the proximity to the diagonal of the perfect predicting, i.e. at all rates of CGPA the prediction was high. Quite to the contrary, the base line model was more dispersion at the low and high CGPA scores. The plots of the residual values showed that the errors of the hybrid model were tightly clustered in both the value of zero and also the distribution of the error was narrow, and this confirms the high level of stability, as well as the small value of variance. In any case, the larger residual distribution in the baseline models signified the predictability of other residuals inconsistency. These sorts of visual traditions are artistic pieces of evidence that will not conflict with the quantitative data but help to conclude that the hybrid model works more best with the hidden data. Visual performance analysis was hence crucial in determining the fact that the perceived improvements could not be limited to the numerical measurements had been acquired but were real predictive reliabilities.

Discussion of Findings:**Figure 7: PCA Results**

The hybrid model is superior in that it has the diversity of strategies to learning interwoven. The model allowed the expansion of features in poly forms to enable the interaction of various aspects of the study (academic and demographic features) to be complex and non-linear. The Principal Component Analysis (PCA) proved useful in removing the redundancy and multicollinearity to allow the model to focus on the most informative elements. Finally the stacking ensemble combined the strength of the Random Forest as well as the SVR and the Decision Tree model transforming it into a stronger and bias and variance resistant model (López-Zambrano et al., 2021). This study has employed reliable and justifiable findings as compared to the false performance inflation employed in the past. The findings demonstrate that learning hybrid framework works better with educational datasets that are quite complex and heterogeneous. Overall, it is demonstrated that the effect of methodological choices on the outcome of generalisation and predictive accuracy was straight positive.

Summary: The presented chapter has disclosed that the proposed PCA based hybrid stacking model is an encouraging, scalable, and realistic predictive model of student academic performance. The research proved that under careful preprocessing and constant evaluation, and comprehensive comparison of the findings with reference models, one might conclude that hybrid learning may produce a significant beneficial impact on the predictive power and stability. The conclusions made that stand alone models though convenient are when it comes to handling complex educational information. In their turn, the hybrid approach does its job well in encompassing the accuracy, generalisation, and interpretability. The excellence of the suggested model was always evidenced with the help of the analogy tests and statistical tests. Overall, the findings give strong empirical reasons regarding the direction to take as hybrid machine learning algorithms in the educational analytics, which brings about the possibility of early intervention systems or data-founded academic decisions.

Reference List

- Alalawi, K., Athauda, R. and Chiong, R., 2023. Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review. *Engineering Reports*, 5(12), p.e12699. <https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/eng2.12699>
- Albreiki, B., Zaki, N. and Alashwal, H., 2021. A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*, 11(9), p.552. <https://www.mdpi.com/2227-7102/11/9/552>
- Alsariera, Y.A., Baashar, Y., Alkaws, G., Mustafa, A., Alkahtani, A.A. and Ali, N.A., 2022. Assessment and evaluation of different machine learning algorithms for predicting student performance. *Computational intelligence and neuroscience*, 2022(1), p.4151487. <https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/4151487>
- Balaji, P., Alelyani, S., Qahmash, A. and Mohana, M., 2021. Contributions of machine learning models towards student academic performance prediction: a systematic review. *Applied Sciences*, 11(21), p.10007. <https://www.mdpi.com/2076-3417/11/21/10007>
- Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H.Y. and Hussain, A., 2023. Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1), pp.905-971. https://www.researchgate.net/profile/Muftah-Baroud-2/publication/373040567_Compressive_strength_in_cement_mortars_via_impulse_excitation_technique_and_genetic_algorithm/links/677dc2f679d6d00f5f5ce8c4/Compressive-strength-in-cement-mortars-via-impulse-excitation-technique-and-genetic-algorithm.pdf
- Issah, I., Appiah, O., Appiahene, P. and Inusah, F., 2023. A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision analytics journal*, 7, p.100204. <https://www.sciencedirect.com/science/article/pii/S2772662223000449>
- López-Zambrano, J., Torralbo, J.A.L. and Romero, C., 2021. Early prediction of student learning performance through data mining: A systematic review. *Psicothema*, 33(3), p.456. <https://www.psicothema.com/pdf/4692.pdf>
- Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J. and Phasinam, K., 2023. Classification and prediction of student performance data using various machine learning algorithms. *Materials today: proceedings*, 80, pp.3782-3785. https://www.researchgate.net/profile/Harikumar-Pallathadka/publication/353611098_Classification_and_prediction_of_student_performance_data_using_various_machine_learning_algorithms/links/610602d4169a1a0103cc9854/Classification-and-prediction-of-student-performance-data-using-various-machine-learning-algorithms.pdf
- Yağcı, M., 2022. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1), p.11. <https://link.springer.com/content/pdf/10.1186/s40561-022-00192-z.pdf>
- Zeineddine, H., Braendle, U. and Farah, A., 2021. Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, 89, p.106903. https://www.researchgate.net/profile/Hassan-Zeineddine-2/publication/346393467_Enhancing_Prediction_of_Student_Success_Automated_Machi

ne_Learning_Approach/links/5fbf805a299bf104cf7a8563/Enhancing-Prediction-of-Student-Success-Automated-Machine-Learning-Approach.pdf

- Algarni, A., 2016. Data Mining in Education. International Journal of Advanced Computer Science and Applications, 7(6), 2016.
DOI:10.14569/IJACSA.2016.070659
- Chahboun S, Maaroufi M, 2021. Principal Component Analysis and Machine Learning Approaches for Photovoltaic Power Prediction: A Comparative Study, Applied Science, 11(17), 7943.
<https://doi.org/10.3390/app11177943>
- Alberto Jimenez-Macias A, Merino PJM, Moreno-Marcos PM, Delgado-Kloos C, 2025. Evaluation of traditional machine learning algorithms for featuring educational exercises, Applied Intelligence, 55(7), 501.
<https://doi.org/10.1007/s10489-025-06386-5>