International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 04 | Apr 2024 www.irjet.net p-ISSN: 2395-0072

Student Performance Analyzer

Athul Raj V M[1], Surya Sushad[2], Sneha S[3] , Rachana Ramachandran R P[4]

[1]Student, Dept.of Information Technology, KMCT College of Engineering, Kerala, India
[2]Student, Dept.of Information Technology, KMCT College of Engineering, Kerala, India
[3]Student, Dept.of Information Technology, KMCT College of Engineering, Kerala, India.
[4]Assistant Professor, Dept.of Information Technology, KMCT College of Engineering, Kerala, India

---

Abstract - *Education plays a crucial role in shaping students' future prospects. Additional assignments and projects assigned by instructors can bolster the academic performance of students struggling academically. However, a significant challenge lies in the early identification of students who are at risk. Researchers are actively exploring this issue using Machine Learning techniques. Machine learning finds applications across various domains, including the early identification of at-risk students and providing them with necessary support from instructors. This research delves into the outcomes achieved through Machine Learning algorithms in identifying at-risk students and mitigating student failure. The primary objective of this project is to develop a hybrid model using ensemble stacking methodology to forecast at-risk students accurately. A range of Machine Learning algorithms such as Naïve Bayes, Random Forest, Decision Tree, K-Nearest Neighbors, Support Vector Machine, AdaBoost Classifier, and Logistic Regression are employed. Each algorithm's performance is assessed using diverse metrics, and the study showcases the hybrid model that amalgamates the most effective algorithms for prediction. The model is trained and tested using a dataset containing students' demographic and academic information. Additionally, a web application is designed to facilitate the efficient utilization of the hybrid model and obtain prediction results. The study reveals that employing stratified k-fold cross-validation and hyper parameter optimization techniques enhances the model's performance. Furthermore, the hybrid ensemble model's efficacy is evaluated using two distinct datasets to underscore the significance of data features. In the first combination, utilizing both demographic and academic data, the hybrid model achieves an accuracy of 94.8%. Conversely, when solely academic data is utilized in the second combination, the accuracy of the hybrid model increases to 98.3%. The study's focal point is early prediction of at-risk students, enabling educators to offer timely assistance to students struggling academically.*

Key Words: Ensembling Stacking Method, Machine Learning Techniques, dropout prediction.

1.INTRODUCTION.

Throughout the academic term, certain students may encounter obstacles in their studies due to various factors, including psychological issues, familial dynamics, peer pressure, or inadequate teacher support. These challenges can significantly impede the academic progress of these students, necessitating timely intervention from educators to identify and offer assistance. Predicting students' academic performance enables teachers to identify those in need of additional courses, supplementary assignments, or support services. However, in educational institutions with a sizable student body, analyzing individual student performance can be daunting.

Early identification of students who might face academic difficulties allows for targeted interventions to improve their prospects before problems escalate. This project primarily focuses on high school students, recognizing the substantial impact of their academic achievements on future educational pursuits. By utilizing a dataset comprising both academic and demographic data, the project aims to assess students' success or failure based on established educational criteria. Success is determined based on year-end average scores, with a passing grade set at 50 or above, and a failing grade below 50. According to educational regulations, students with a year-end general average grade below 50 may progress to the next grade provided they have no more than three failed courses. Identifying potentially struggling students early in the academic year is crucial, though challenging due to the large student population and resource constraints.

To address this challenge, various techniques are employed to identify at-risk students, acknowledging the complexities educators face in this endeavor. The project emphasizes the importance of employing differentiated techniques to identify students facing precarious academic situations, particularly in high school settings. A departure from traditional methods, the project aims to develop a hybrid model using machine learning techniques. This hybrid model integrates supervised learning algorithms, including Naive Bayes, Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbor, Logistic Regression, and AdaBoost, with the goal of achieving superior predictive outcomes. These algorithms' accuracy is evaluated using diverse metrics to identify the top performers. Ensemble methods such as Bagging, Boosting, and Stacking are employed to combine multiple machine learning techniques, with the stacking ensemble learning approach selected for creating the hybrid model.

The project comprises two phases: Phase 1 involves determining the base model with the best prediction results from the initial models, while Phase 2 focuses on selecting the meta-classification model and optimizing the combination of predictions from the base models. Stratified k-fold cross-validation of the base models is used to effectively prepare the training dataset for the meta model. The primary objective of the project is to proactively identify high school students at risk of academic failure and provide targeted support to enhance overall student success. The research utilizes a newly collected dataset from high school students, including demographic information and academic performance indicators gathered through questionnaires. The study highlights the importance of academic data in identifying at-risk students and introduces a hybrid ensemble model to improve prediction accuracy.

2. LITERATURE SURVEY.

2.1 Early Predicting of Students Performance in Higher Education.

Evaluating students' learning performance stands as a pivotal aspect in assessing the effectiveness of any educational system. It plays a critical role in addressing challenges within the learning process and serves as a key metric for measuring learning outcomes. The utilization of data-driven insights to enhance educational systems has given rise to the field of educational data mining (EDM). EDM involves the development of methodologies to analyze data collected from educational environments, facilitating a deeper and more precise understanding of students and thereby enhancing educational outcomes.

In recent years, there has been a significant surge in the adoption of machine learning (ML) technology. Researchers and educators leverage the insights derived from data mining in education, including metrics related to success, failure, and dropout rates, to forecast and simulate educational processes. Thus, this study undertakes an examination of students' performance using data mining techniques. It employs both clustering and classification methodologies to ascertain the early-stage impact on GPA.

In employing the clustering technique, the study utilizes the T-SNE algorithm for dimensionality reduction, considering factors such as admission scores, initial level courses, academic achievement tests (AAT), and general aptitude tests (GAT) to explore their relationship with GPA. Regarding the classification approach, the study conducts experiments using various machine learning models to predict student performance in early stages, incorporating features such as course grades and admission test scores. Additionally, diverse assessment metrics are employed to evaluate the models' efficacy. The findings suggest that educational systems can proactively mitigate the risks ofstudent failure during the early stages of their academic journey.

2.2 Predicting Students' Academic Performance Using Artificial Neural Network.

In today's contemporary society, technology pervades nearly every aspect of human life, including the educational sector. Among the core objectives of educational institutions is the enhancement of students' academic performance. Traditionally, many secondary schools have relied on manual methods such as assumptions and mock exams to gauge students' progress and strive for improvement. However, this approach, employed over the years, has proven to be ineffective and inefficient, evident from the consistent decline in WAEC results among students in both private and public secondary schools in Nigeria. Despite the presence of Information and Communication Technology (ICT) and student data, most schools have not tapped into the potential of this data to address the issue at hand.

Given the limitations of existing approaches in improving students' academic performance, there is a pressing need for a more effective strategy. Hence, this research proposes an approach that harnesses AI technology to achieve better outcomes. The study focuses on leveraging ICT and Artificial Neural Networks (ANN) to predict students' academic performance. Through a combination of qualitative and quantitative methodologies, the requirements for such a system were identified. WEKA was utilized as the simulation environment to evaluate the proposed algorithm based on the identified factors. The test results from the prototype indicate that the model holds promise for predicting students' academic performance in secondary schools.

The research introduces concepts such as predictive analytics, predictive modeling, and Artificial Neural Networks, with WEKA serving as the platform for analyzing academic performance data in conjunction with contextual factors. The methodology encompasses a mixed-method research design, incorporating both qualitative and quantitative data analysis techniques. The prototype's abstraction is based on fundamental elements such as the sigmoid function, which underpins the neural network's predictive capabilities. This research endeavors to bridge the gap between traditional assessment methods and innovative AI-driven approaches to enhance students' educational outcomes.

2.3 Using Data Mining To Predict Secondary School Student Performance.

Despite improvements in the educational attainment of the Portuguese population over recent decades, Portugal still lags behind in Europe due to its persistently high rates of student failure. Particularly concerning is the lack of success

in foundational subjects like Mathematics and the Portuguese language. Conversely, in the realm of Business Intelligence (BI) and Data Mining (DM), which focus on extracting valuable insights from raw data, there exist promising automated tools that could benefit the education sector. This study aims to address student performance in secondary education by employing BI/DM techniques. Real-world data, including student grades, demographic information, social factors, and school-related attributes, was gathered from school reports and questionnaires. The study focused on modeling the two core subjects, Mathematics and Portuguese, through binary/five-level classification and regression tasks.

Four DM models - Decision Trees, Random Forest, Neural Networks, and Support Vector Machines were employed, along with three input selection strategies, including the inclusion/exclusion of previous grades. The results indicate that achieving a high predictive accuracy is feasible, particularly when grades from the first and/or second school periods are available. While past academic performance significantly influences student achievement, explanatory analysis revealed other pertinent factors, such as the number of absences, parental occupation and education level, and alcohol consumption. As a direct outcome of this research, the development of more effective student prediction tools is envisioned, aiming to enhance the quality of education and optimize school resource management.

This study underscores the potential of BI/DM techniques to address persistent challenges in student achievement, offering insights that could inform targeted interventions and support mechanisms in the education sector.

3. PROBLEM STATEMENT.

3.1 Existing System

The pressing concern at the heart of this project is the systemic hurdle within modern education systems. Traditional methods of identifying academically vulnerable students exhibit inherent limitations, prompting a necessary shift towards integrating machine learning. This shift aims to proactively address academic challenges comprehensively. The deficiencies in conventional assessment methods, coupled with subjective teacher observations, lead to delayed identification of struggling students. Consequently, there's a heightened risk of failure and a persistent gap in providing timely interventions.

Hence, the core objective of this initiative is to leverage machine learning's advanced analytics to meticulously analyze diverse datasets. These datasets encompass academic records, attendance, behavior, and socio-economic factors. The ultimate goal is to develop a sophisticated predictive model capable of identifying subtle patterns indicative of at-risk students early in their academic journey.

4. PROPOSED SYSTEM

4.1 Introduction

The proposed system is to identify high school students at risk before the end of the education period and to support the education of high school students. The purpose of our system is to increase the success performance of students, as well as to identify students who may fail in the class before the end of the semester and to provide timely support by the teacher to such students. Teachers can be informed about students at risk (students who may fail) as determined by the hybrid model and additional study material may be provided to these students by the teachers. By analyzing the characteristics of the data set obtained from the students, the characteristics that affect the school performance of the students can be determined. Thus, more efficient results can be obtained by using these data while creating a hybrid model.

4.2 Methodology

4.2.1 Data Collection

Utilizing a dataset sourced from high school students, the collection process involved the distribution of a structured questionnaire. Tailored to accommodate diverse backgrounds, social contexts, and talents, the questionnaire comprised two sections: one focusing on demographic details and the other on academic aspects. Responses varied from binary choices like Yes/No to textual inputs such as numbers or words. Google Forms facilitated the survey creation.

The gathered dataset aligns with the project's objectives, encompassing crucial features like study habits, exam and homework scores, future educational aspirations, and extracurricular engagements. Acknowledging the diversity among students, the data collection emphasized their distinct characteristics. This dataset significantly contributes to the project by incorporating contemporary information, enabling the development of a hybrid model. By addressing prevailing student issues and considering educational system dynamics, the dataset enhances the model's efficacy in identifying at-risk students.

4.2.2 Pre-processing and Feature Engineering

To prepare the dataset for analysis, it undergoes a crucial step called data preprocessing. This process involves transforming raw data into a more interpretable format suitable for model utilization. Detecting missing, inconsistent, outlier, and erroneous data is essential to avoid inaccurate estimation outcomes due to incomplete or inconsistent data. In line with the project's objectives, a "Pass" column was introduced to monitor students' course success. Success is defined as achieving an average of at least 50 in both semesters combined. Accordingly, students with a year-end average of

50 or higher are marked as successful (assigned the binary value '1'), while those below 50 are classified as unsuccessful (assigned '0'). This addition of the "Pass" feature serves as a predictive target for the models.

Following data preprocessing, visualization techniques were employed using various graphs to comprehend data feature relationships and assess their impact on student success performance. Data visualization, facilitated by the Seaborn library in Python, is instrumental in gaining insights into dataset characteristics. This analysis aims to identify key data attributes crucial for model training, with the results and interpretations of the visualizations forming part of the evaluation process.

4.2.3 Building the model for initial results

The initial phase involved employing various supervised machine learning algorithms such as Random Forest, K-NN, Decision Tree, SVM, Naive Bayes, Logistic Regression, and AdaBoost to generate preliminary model results. These selections were based on a comparative analysis of algorithms utilized in prior literature to identify those demonstrating superior performance, which were then chosen for this study. Before constructing the hybrid model, each algorithm underwent individual evaluation and comparison, utilizing both academic and demographic data from the dataset. Additionally, separate assessments were conducted using only academic data to assess its impact on prediction outcomes. The target label for the models was set as the "Pass" column, and default hyper parameters were utilized for all algorithms.

Upon reviewing the initial results, adjustments to the hyper parameter values of algorithms exhibiting the most promising predictions were considered before creating the hybrid model. The proposed approach employed the stratified k-fold cross-validation method to partition the dataset into training and validation folds. Rather than traditional train-test splitting, this method facilitated training and testing models with each data feature individually, thereby generating prediction outcomes. Utilizing cross-validation mitigated potential issues arising from model reliance on specific data features during training, thus addressing concerns of overfitting and assessing the model's performance consistency. Set with a k value of 10, the stratified K-Fold Cross Validation ensured representative sampling across folds, contributing to more robust prediction performance. Machine learning algorithms used to obtain the first model results involves:

- Logistic Regression:

  Logistic regression does binary classification. Project case, the target feature is the ‘‘pass’’ column. Here, there are two possibilities, (1) student will be successful (represented by 1) and (2) student will fail (represented by 0). The logistic regression function is given below. In Equation ,(i) p is the probability of the target event, (ii) the set {x1, x2, . . . xn} represents the independent variables, (iii) the set {b1, b2, . . . , bn} represents coefficients of the logistic regression, (iv) b0 represents the bias (or intercept) term. The output value will be modeled as binary value (1 or 0).

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + \ldots + b_nx_n$$

$$p = \frac{e^{(b_0+b_1x_1+b_2x_2+\ldots+b_nx_n)}}{1+e^{(b_0+b_1x_1+b_2x_2+\ldots+b_nx_n)}}$$

- K-nearest Neighbors:

  After hyperparameter adjustments, used Euclidean as the distance metric parameter for KNN algorithm. Euclidean distance formula as follows.

$$d(x,y) = \sqrt{(x_1-y_1)^2+\ldots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

  According to the Euclidean distance, x and y represent the two points in the coordinate plane. Consequently, K-Nearest Neighbors as follows. In Equation , (i) X is dataset features, (ii) k is the number of neighbors, (iii) Y is target class, (iv) R represents the set of observations of the k nearest points, (v) indicator variable as I(yi = j).

$$P(Y=j|X=x) = \frac{1}{k}\sum_{where\ i\ \in R} I(y_i=j)$$

- Decision Tree:

  This algorithm is one of the supervised learning algorithms. It is a model that use to classify students according to their school performance status (successful or unsuccessful). After making the parameter adjustments, entropy criterion was used as parameter. Entropy is the impurity measurement and calculated based on feature Y. In equation , X represents the target variable (pass column), Y represents the feature in the dataset and pi represents the probability of target i at the node. It can be mathematically expressed as follows.

$$E(X,Y) = \sum_{i=1}^{n} -p_i log_2(p_i)$$

- Naive Bayes:

  Use Bayes theorem along with the conditional probability formula to calculate the probability of an event to occur. In project case, this event can represent whether students are successful or not. In equation , X is vector of features and Y is class variable. The way this algorithm works is by calculating the probability of each event for a variable and classify that variable according to the

highest probability outcome. The following expression of Bayes theorem calculates the probability of event Y occurring when event X occurs.

$$P(Y \mid X) = \frac{P(X \mid Y)\, P(Y)}{P(X)}$$

P(Y|X) and P(X|Y) = Conditional probability.
P(Y) = Prior probability of the class variable.
P(X) = Prior probability of the predictor.

- Random Forest:

Depending on the features in the data set, classification is conducted to predict whether the student will be successful or not. The random forest algorithm classifies by employing a combination of multiple decision trees and chooses the decision tree that gives the best outcome as the prediction result. The entropy function (Equation) was chosen to measure the probability of a specific outcome. Furthermore, the random forest model for our problem maybe expressed as follows.

$$D = \begin{bmatrix} x_{1,1} & x_{2,1} & \dots & x_{m,1} & y_1 \\ \dots & \dots & \dots & \dots & \dots \\ x_1, n & x_{2,n} & \dots & x_{m,n} & y_n \end{bmatrix}$$

Here, (i) D represents the training dataset, (ii) the variables {x1, x2, . . . , xm} represent data features in the dataset, (iii) There are n samples of each feature, (iv) the variables {y1, . . . , yn} represent the class (target) label. From this set, Can get sample random subsets including data samples. The random forest algorithm performs decision tree combinations with each created subset. As a result, the best output is considered based on the classification outcome.

- AdaBoost:

The AdaBoost classifier was used as another machine learning algorithm. This algorithm combines the weak learners with importance weights to get a strong learner. Each time, the training data is updated based on the data points and it is used in the next learner model. In Equation , where (i) α is the classifier coefficient (applied weight), (ii) M is weak classifiers, (iii) ym(x) is weak classifier outputs.

$$Y(x) = sign\left(\sum_{m=1}^{M} \alpha_m y_m(x)\right)$$
$$where\ \alpha_m = \frac{1}{2} \ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$$

- Support Vector Machine:

SVM is a supervised learning algorithm which can be used to solve the classification problems. To conduct the classification, a linear line is drawn that separates the sample classes and then the optimal hyperplane is found. It classifies the sample data points by deciding which class they belong to. The SVM model can be expressed mathematically as in Equation , where (i) f(x) is the main function, (ii) (x) is the feature map, (iii) b is the bias term, (iv) w is the weight vector. In Equation 11, function y can be defined according to the f(x) function. If the sample point is under the hyperplane, it is evaluated as -1. If the sample point is above or directly on the hyperplane, it is evaluated as 1.

$$f(x) = w^T \Phi(x) + b$$
$$y = \begin{cases} 1\ if\ f(x) \geq 0 \\ -1\ if\ f(x) < 0 \end{cases}$$

4.2.4 Creating the hybrid ensemble model

More than one model is created to determine the predictive model that provides the best accuracy performance. The accuracy performance of each model is different due to the errors made by the models considering different points in the data set. Ensemble learning technique is a good way to use to improve the performance of the models. With Ensemble learning, results are combined using multiple best performing models. Thus, a clearer and higher estimation result can be obtained.

The stacking method has been used in this project. It is one of the ensemble learning techniques, to create the hybrid model. With this approach, the performance of the predictive model is increased, and margins of error are reduced. The hybrid model was created with the models used for initial results by the stacking method. Each algorithm was tested for metalearner, and the algorithm that gave the best performance result was used as a meta-learner. According to the results obtained, Naive Bayes, Random Forest, Decision Tree, AdaBoost, Logistic Regression and KNN algorithms were used as base learners. The Support Vector machine algorithm was used as a meta learner.

The general procedure of the stacking method proceeds as follows. First-level learning algorithms are represented as base learner models, and the second-level learning algorithm is represented as meta-learner model. These learners are combined to create the stacking model. First, base learners are trained with the training part of the data set, and to train the meta learner, it is necessary to create a different training set than the data set used to train the base learners . The results of the predictions are obtained by testing the base learners with the test set. The prediction outputs obtained from the base learners are used as the input of the meta learner. Final prediction results are obtained after training the meta learner with the new created data set. The method presented in this project, stacking 10-fold cross validation is used to create a new training dataset for meta learner. Project include, SVM was used as the meta learner model and as a result of the meta learner model, the prediction results of the final hybrid modelwere obtained.

The stacking ensemble method was proposed for predicting at-risk students. First, base learner models were trained with the part of the data set created for training. Subsequently, a new training set is created by the model based on the prediction results obtained from the base learner models. The original data labels are still considered the target class when creating the new dataset. The following first expression, D is the data set; it includes data features. second expression is the generated new data set which is expressed as D'. This new generated data set will be used for meta model training. D' includes predictions of base learners and features. Finally, after training the meta learner, we can get predictions of the ensemble model.

$$D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$$
$$D' = \left\{\left(z_{i,1}, \ldots, z_{i,n}\right), y_i\right\} \text{ when } i = 1, 2, \ldots, n$$

4.2.5 Performance Evaluation

After the final hyper parameter adjustments were made in the hybrid model, the performance of the model was measured with various metrics. There are several metrics for evaluating machine learning models. Since classification models were used in this case, the performance of the hybrid model was evaluated using accuracy, recall, precision, AUC-ROC curve and F1 score metrics. The formulas of the accuracy, recall, and precision metrics are given in Equation.

$$\text{Recall} = \frac{\text{True Positive}}{\textit{True Positive} + \textit{False Negative}}$$

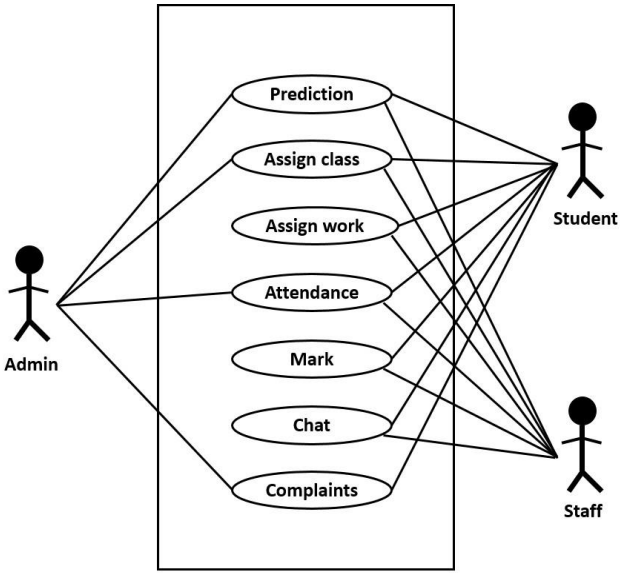
$$\textit{Precision} = \frac{\text{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

$$\textit{Accuracy} = bb\frac{\textit{True Positive} + \textit{True Negative}}{\textit{Total number of samples predicted}}$$

5. SYSTEM DESIGN AND DEVELOPMENT

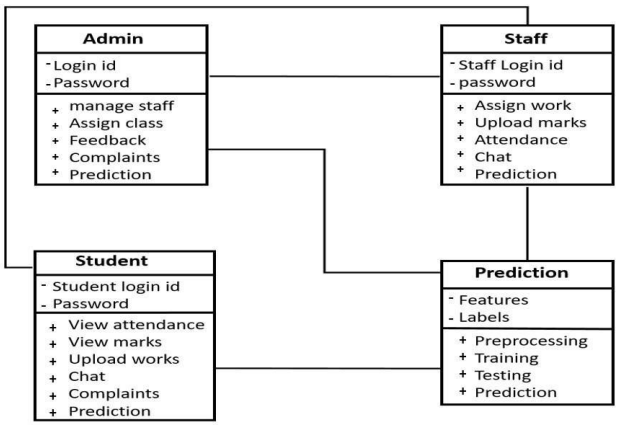
5.1 Use Case Diagram

A UML use case diagram is the primary form of system/software requirements for a new software program under development. Use cases specify the expected behavior, and not the exact method of making it happen. It summarizes some of the relationships between use cases, actors, and systems. It is an effective technique for communicating system behavior in the user's terms by specifying all externally visible system behavior. Here three actors are Admin , Staff and student.



5.2 Class Diagram

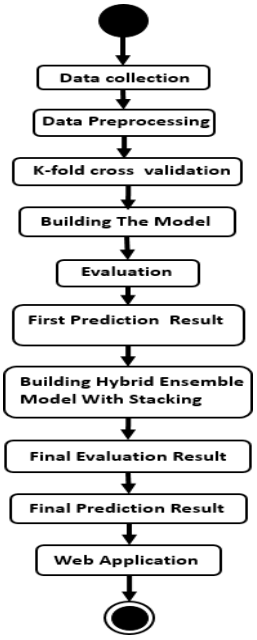
A class diagram is a static representation illustrating the attributes, operations, and constraints of a class within a system. It serves as a fundamental tool in modeling object-oriented systems due to its direct correlation with object-oriented languages. Comprising classes, interfaces, associations, collaborations, and constraints, it offers a comprehensive view of the system's structure. Class diagrams are also recognized as structural diagrams within the realm of UML modeling.



5.3 Activity diagram

Activity diagrams serve as visual depictions of stepwise workflows encompassing activities, actions, and decision points, accommodating choice, iteration, and concurrency. Within the Unified Modeling Language (UML), they aim to depict computational and organizational processes, along with data flows intersecting these activities. While primarily illustrating control flow, activity diagrams can

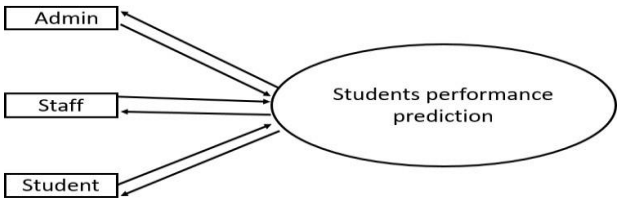
also incorporate elements representing the flow of data stores.



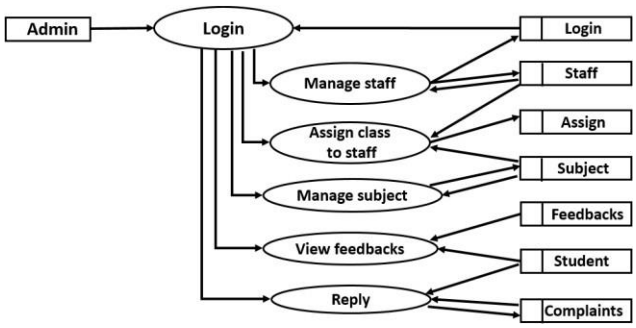
5.4 Data Flow Diagram

A data-flow diagram serves as a visual representation illustrating the movement of data within a process or system, typically an information system. It outlines the inputs and outputs of each entity, as well as the process itself. Unlike control flow diagrams, data-flow diagrams lack decision rules and loops. Instead, flowcharts are typically used to represent specific operations based on the data.

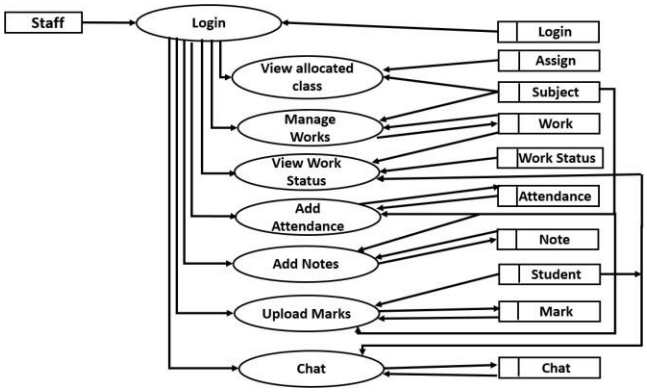
LEVEL 0



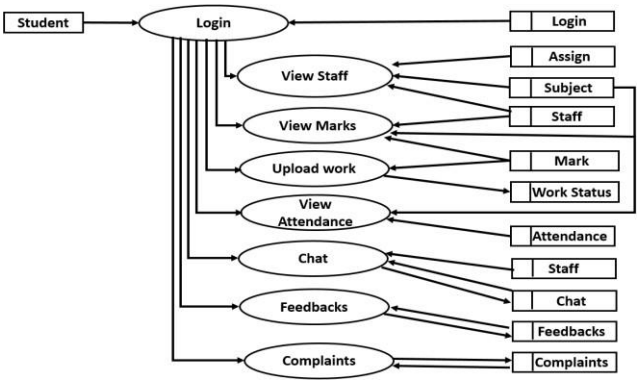
LEVEL 1.1



LEVEL 1.2



LEVEL 1.3



6. RESULTS AND DISCUSSION

The predictive outcomes aid teachers in pinpointing students at risk and mitigating potential failures. Initially, individual model evaluations yielded performance results. It was noted that employing the Stratified K-Fold Cross Validation method positively impacted model performance. Additionally, optimizing hyperparameters was found to enhance model effectiveness. Hence, utilizing stratified 10-fold CV and conducting hyperparameter optimization were deemed to improve machine learning model performance. Among the individually evaluated models, Logistic Regression exhibited the highest accuracy at 94.4%.

Following the proposed approach, a hybrid model was constructed using the stacking method, experimenting with various supervised algorithms as meta learners. Optimal performance for the hybrid model was achieved when SVM served as the meta learner, yielding an accuracy of 94.8% and a precision of 96.8%. Comparing hybrid models created with different meta learners revealed that SVM yielded the best performance. Performance metrics such as accuracy, precision, recall, F1 score, and AUC score were employed to assess hybrid model performance. Furthermore, a comparison was made between hybrid models utilizing solely academic data versus both demographic and academic

data. Results indicated that while the hybrid model's performance reached 94.8% when utilizing both data types, focusing solely on academic data led to an increased performance of 98.4%. This underscores the significance of academic data in predicting at-risk students. Overall, the hybrid model demonstrated superior performance compared to individual models, as observed from the results obtained.

7. CONCLUSION

To enhance future academic success, it's crucial for teachers to identify students who may be at risk of failing early on. By pinpointing these students ahead of time, teachers can offer additional support and interventions, thus improving their chances of success. The primary goal of this project is to predict students at risk before the end of the academic term, and machine learning techniques are proposed as a solution. Specifically, a hybrid model utilizing supervised machine learning algorithms is developed to address this challenge. Various machine learning algorithms are applied to datasets, and individual models are evaluated. However, a unique approach is taken in creating a hybrid ensemble model using a stacking technique to predict at-risk students. The dataset containing information about high school students, including both demographic and academic attributes, is obtained through a form.

Furthermore, a web application is developed to enable teachers to utilize the hybrid model. By inputting relevant student data into the application, teachers can access information about students' academic performance, thereby identifying those at risk and providing them with necessary assistance. This aligns with the project's objective of creating a hybrid ensemble model to identify at-risk students using the stacking method.

REFERENCES

[1] P. Cortez and A. Silva, ‘‘Using data mining to predict secondary school student performance,’’ in Proc. 15th Eur. Concurrency Eng. Conf. (ECEC), 5th Future Bus.Technol. Conf. (FUBUTEC), 2008, pp. 5–12.

[2] E. Er, ‘‘Identifying at-risk students using machine learning techniques: A case study with IS 100,’’ Int. J. Mach. Learn. Comput., vol. 2, no. 4, pp. 476–480, 2012, doi:10.7763/ijmlc.2012.v2.171.

[3] S. Isljamovıc and M. Suknovıc, ‘‘Predicting students’ academic performance using artificial neural network: A case study from faculty of organizational sciences,’’ Eurasia Proc. Educ. Social Sci., vol. 1, pp. 68–72, May 2014.

[4] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, ‘‘Dropout prediction in e-learning courses through the combination of machine learning techniques,’’ Comput. Educ., vol. 53, no. 3, pp. 950–965, 2009.

[5] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison,‘‘A machine learning framework to identify students at risk of adverse academic outcomes,’’ in Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug.2015, pp. 1909–1918.

[6] H. Agrawal and H. Mavani, ‘‘Student performance prediction using machine learning,’’Int. J. Eng. Res., vol. 4, no. 3, pp. 111–113, Mar. 2015.

[7] L. A. B. Macarini, C. Cechinel, M. F. B. Machado, V. F. C. Ramos, and R. Munoz, ‘‘Predicting students success in blended learning—Evaluating different interactions inside learning management systems,’’ Appl. Sci., vol. 9, no. 24, p. 5523, Dec. 2019.

[8] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala,‘‘Predicting student performance using personalized analytics,’’ Computer.

[9] I. E. Livieris, K. Drakopoulou, V. T. Tampakas, T. A. Mikropoulos, and P. Pintelas,‘‘Predicting secondary school students’ performance utilizing a semi-supervised learning approach,’’J. Educ. Comput. Res., vol. 57, no. 2, pp. 448–470, Apr. 2019.

[10] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, ‘‘Predicting at-risk students at different percentages of course length for early intervention using machine learning models,’’ IEEE Access, vol. 9, pp. 7519–7539, 2021.