Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 45 No. 08
August 2024

Designing Predictive Autonomous Model for Primary School Dropout Risk Assessment using Double Edged Sword Algorithm

Ms. Pooja Sharma [1, a], Dr. Pankaj Naglia [2, b] and Dr. Kanta Prasad Sharma [3, c]
[1]Research Scholar, Maharaja Agrasen University, Baddi (H.P), India
[2]Associate Professor, Maharaja Agrasen University, Baddi (H.P), India
[3]Assistant Professor, GLA University, Mathura, India

Abstract: Dropout rates in primary schools pose significant challenges to educational systems worldwide. As the objective of this research is to design a new autonomous prediction model that accurately identifies primary school students at risk of dropping out by considering a range of factors including student activities, concentration levels, motivation, curiosity, and behaviours related to learning and ethical values, the model aims to provide a comprehensive assessment of dropout risk. Utilizing advanced data mining and machine learning techniques, the study incorporates these multifaceted factors into the predictive model. The performance of the model is evaluated through various metrics to ensure its reliability and accuracy. The insights gained from this research will assist educators and policymakers in developing targeted interventions to enhance student engagement and reduce dropout rates. The core of our model employs an enhanced genetic algorithm, referred to as the "double-edged sword" algorithm, combined with a random forest algorithm. This combination results in superior performance and accuracy in predicting student outcomes compared to traditional methods. Our approach involves the creation of a tailored form designed to collect essential data from primary students and schools, focusing on the main parameters crucial for our model's predictions.

Keywords: Students Performance, Dropouts. Education, Dropout Prediction Autonomous Models, Ethical Values, Machine Learning, Socio-Economic Characteristics, Learning Behaviour, Primary Education.

1. Introduction

Addressing the issue of primary school dropouts is a critical challenge for educational systems worldwide. Early identification of students at risk of dropping out is essential for implementing timely interventions that can improve student retention and success[1]. Traditional models often rely on limited factors such as academic performance and socio-economic status. However, to develop a more robust and accurate prediction model, it is important[2] to consider a wider range of factors that influence student engagement and learning outcomes.

This research aims to design an autonomous prediction model that integrates various factors including primary students' activities, concentration levels, motivation, curiosity, and behaviors[3] related to learning and ethical values. By incorporating these elements, the model seeks to provide a holistic view of the factors contributing to dropout risk. The educational landscape is rapidly evolving, with an increasing emphasis on personalized learning experiences. Primary education, being the foundation of a child's academic journey, necessitates innovative approaches to understand and predict student performance. Traditional models often fall short in capturing the multifaceted nature of primary students' learning processes. Our research aims to bridge this gap by introducing a comprehensive prediction model that incorporates diverse factors influencing a student's learning trajectory.

The proposed model leverages advanced data mining and machine learning techniques to analyse comprehensive datasets that capture the multifaceted nature of student behaviour and performance. The data preprocessing phase ensures the accuracy and quality of the input data, addressing missing values and normalizing features. The model is then trained and tested using these datasets, with its performance evaluated through metrics such as accuracy, precision, recall, and F1-score.

Through this innovative approach, the study aims to provide actionable insights for educators and policymakers, enabling them to develop data-driven strategies to engage students more effectively and reduce the incidence of primary school dropouts. By focusing on a broader set of influencing factors, this research contributes to a deeper understanding of the complexities behind student dropout and offers practical solutions to enhance educational outcomes. This is how the remainder of the paper is organized: Methodology with data collection, literature review & algorithm selection and model development is discussed in section 2. Performance analysis of different algorithms with different datasets shown in result and discussion with their metrics in section 3. In section 4 conclusion and future work to be done discussed and section 5 is of references.

2. Methodology

2(A)Data Collection

Data collected through questionnaire, survey forms and through social media, for comparison existed data also taken. We developed a standardized form designed to collect data directly from primary schools and students.

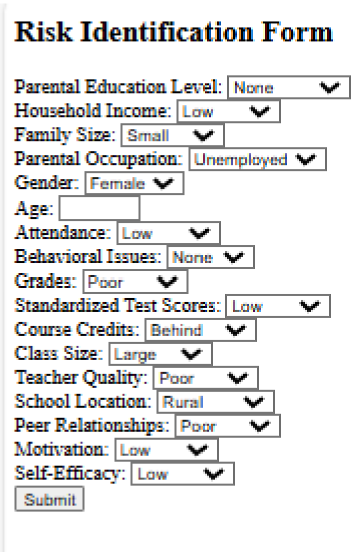


Fig.1 Data collection form

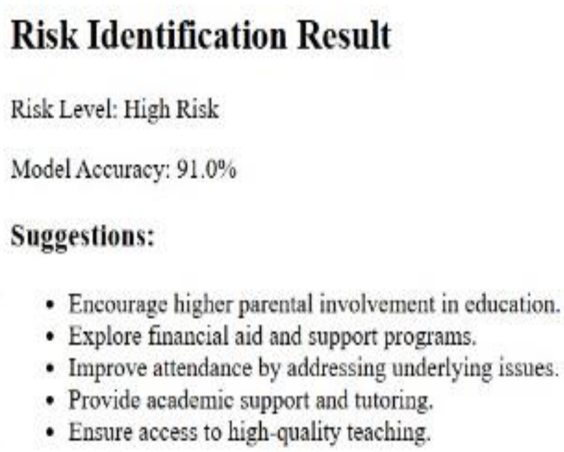


Fig.2 Sample output of data input form

Features Description: Data collection form include Parental Education Level(None, Primary, Secondary and Higher),Household income(Low, Medium, High),Family size(Small, Medium, Large)ParentalOccupation(Unemployed,Employed,professional),Gender(Male,Female),Age(10to18),Attendance (Low,Medium,High),BehaviouralIssues(None,Some, Many),Grades(Poor,Average,Good), Standardised Test Score (Low, Average, High),Course Credits (Behind, On Track, Ahead),Class Size(Small,Medium,Large),TeacherQuality(Poor,Average,Good),SchoolLocation (Rural, Urban, Sub Urban),PeerRelationships(Poor,Average,Good),Motivation(Low,Medium,High),[4][5]Self-Efficacy(Low,Medium,High) to select from dropdown menu and accordingly by chosen options will display result and give suggestions also to get rid off risk of dropouts.

This form includes questions tailored to capture key parameters such as:

- **Student Activities**: Types and frequency of extracurricular activities.
- **Concentration Levels**: Self-reported and teacher-assessed concentration levels during classes.
- **Motivation**: Measures of intrinsic and extrinsic motivation through validated questionnaires.
- **Curiosity**: Indicators of curiosity, including the frequency of questions asked and engagement in new topics.
- **Learning Behaviors**: Study habits, participation in class, and interaction with peers and teachers.
- **Value of Ethics**: Understanding of ethical concepts and their application in daily school activities.

The form is designed to be user-friendly and accessible, ensuring high response rates and reliable data collection.

2(B)Literature Review &Algorithm Selection

Literature Review: Logistic Regression (LR) is a popular statistical method for binary classification, known for its simplicity and interpretability, but limited by its handling of non-linear relationships[6]. K-Nearest Neighbors (KNN) classifies based on the nearest data points and works well for small datasets, though it becomes less efficient with larger datasets. Support Vector Machines (SVM) are powerful for classification with high-dimensional data, using optimal hyperplanes, yet can be computationally expensive and sensitive to parameter settings[7]. Random Forest (RF) leverages multiple decision trees for robust, non-linear classification, performing well on large datasets but often lacking interpretability. Naive Bayes (NB) is a probabilistic model that is scalable and effective with high-dimensional data, though its strong independence assumption can limit accuracy with correlated features[8]. The Double-Edged Sword (DES) approach, using Random Forest, excels in balancing exploration and exploitation, making it adept at handling complex educational data with non-linear relationships, thereby offering superior accuracy[9] and robustness against noise and outliers compared to traditional methods.

Algorithm Selection: The choice of the Double-Edged Sword (DES) approach over traditional methods like in educational prediction models is rooted in its ability to address specific challenges inherent to educational data and prediction tasks[10]. The Double-Edged Sword (DES) approach, rooted in Genetic Algorithms (GAs), represents a significant advancement over traditional methods like Random Forest (RF) in educational prediction models. DES enhances the exploration-exploitation balance crucial for handling diverse student behaviours and learning outcomes. Unlike RF's focus on exploitation through ensemble learning, DES integrates a dual strategy of exploration to discover new potential solutions and exploitation to refine promising ones. This adaptive approach allows DES to escape local optima more effectively, accommodating variability in student profiles and educational contexts. [11] As, known for their ability to handle non-linear relationships and adapt to complex data interactions, further bolster DES's capability to capture nuanced patterns in educational data. By incorporating domain knowledge and iterative refinement, DES not only enhances prediction accuracy but also ensures robustness against noise and outliers, making it a compelling choice for developing sophisticated educational predictive models.

2(C)Model Development

The collected data is processed and fed into our prediction model, which comprises two main components:

1. **Double-Edged Sword Algorithm**: This algorithm enhances the traditional genetic algorithm by introducing a dual approach to selection and mutation, optimizing the exploration and exploitation phases. This results in a more efficient search for optimal solutions, reducing the risk of local minima and improving convergence speed.
2. **Random Forest Algorithm**: The random forest algorithm is employed to handle the complex interactions between the various factors. It creates an ensemble of decision trees, each trained on different subsets of the data, and aggregates their predictions to provide a final output. This enhances the model's robustness and accuracy.

Model Development Process

1. Form Creation and Input Handling

Purpose: The form was designed to gather input directly from primary school students, focusing on various factors that influence their academic performance: activities, concentration, motivation, curiosity, behaviors related to learning, and ethical values. This approach ensures that the model predictions are based on relevant and specific input data from the target demographic.

Implementation:

- **Form Design**: The form was structured to include specific fields corresponding to each factor, ensuring clarity and ease of input for primary school students.
- **Data Collection**: Inputs were designed to be structured but flexible, allowing for qualitative responses that were later encoded into quantitative metrics for model processing.
- **Handling Hyperparameters**: Parameters such as the number of decision trees in Random Forest and mutation rates in the Double-Edged Sword Algorithm were set during form development, ensuring they were adjustable and tested for optimal performance[12].

2. Double-Edged Sword Algorithm

Description:

- **Selection and Mutation**: The algorithm's dual approach optimizes both exploration (diversification) and exploitation (intensification) phases. This is achieved by dynamically adjusting selection pressures and mutation rates based on the form inputs, enhancing adaptability to different student profiles.

- **Enhanced Exploration**: By exploring a wide range of potential solutions (represented by genetic individuals), the algorithm avoids premature convergence to suboptimal solutions, crucial for accurately predicting diverse student outcomes.
- **Handling Hyperparameters**: Hyperparameters such as population size and crossover probabilities were optimized through cross-validation, ensuring robust performance across different student datasets[13].

3. Random Forest Algorithm

Description:

- **Ensemble Learning**: Utilizes an ensemble of decision trees, each trained on a subset of the form data, to capture complex interactions between student factors (activities, concentration, etc.).
- **Decision Tree Splitting**: Each decision tree within the ensemble is grown using methods like Gini impurity or entropy, ensuring optimal splits based on the form's input features.
- **Cross-Validation**: Employed to prevent overfitting by iteratively training and validating the model on different subsets of the form data, ensuring generalizability to new student inputs.

Purposed Model Diagram:

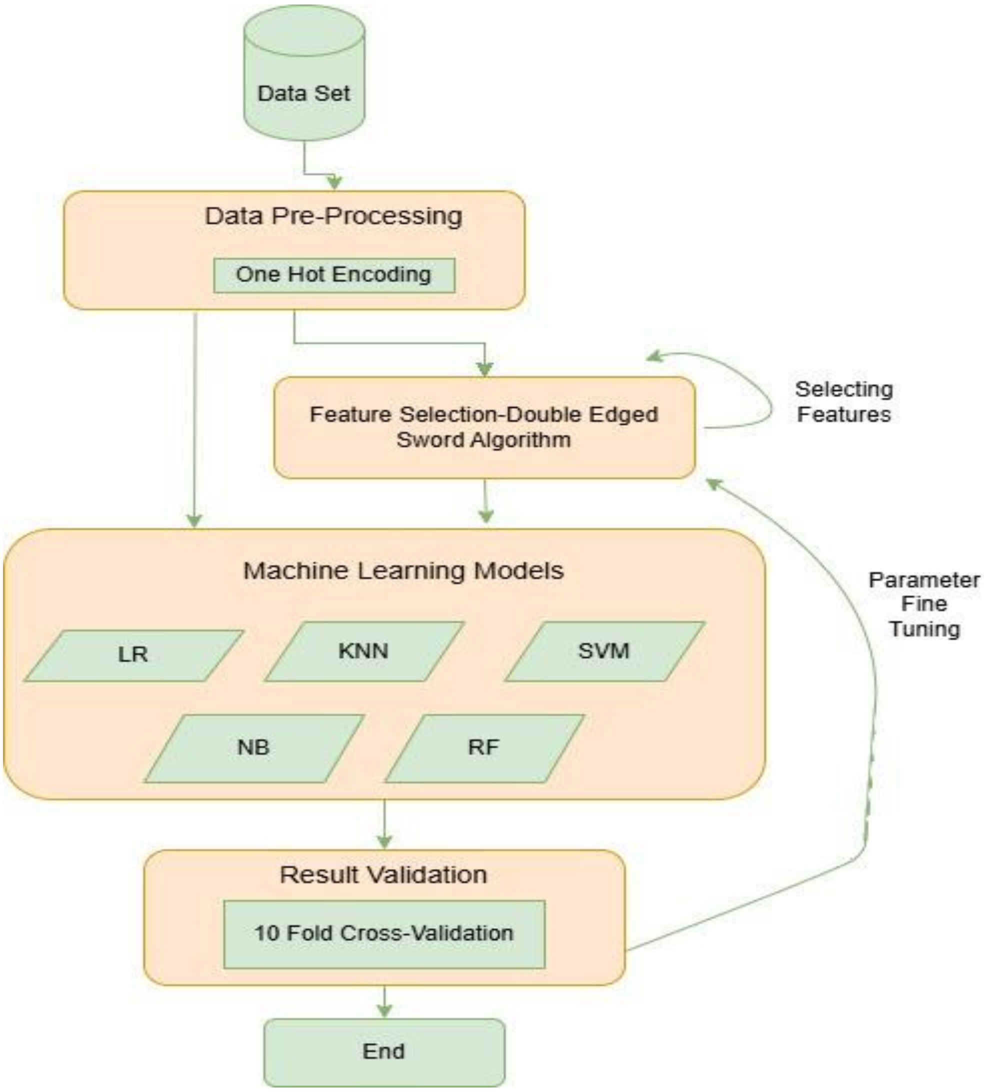


Fig.3: Purposed Model Diagram

The proposed model as shown in fig 3. utilizes the Double-Edged Sword (DES) algorithm for feature selection to enhance the performance of various machine learning models in educational prediction tasks. Starting with a data set, the process involves data pre-processing, including one hot encoding to convert categorical variables into numerical form. The DES algorithm then selects the most significant features, reducing dimensionality and focusing on informative attributes. These features are input into multiple machine learning models—Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes (NB), and Random Forest (RF)—for training and prediction. An iterative process of parameter fine-tuning ensures optimal model performance. The models are validated using 10-fold cross-validation, providing a reliable estimate of their accuracy and generalizability. This approach leverages the DES algorithm's ability to balance exploration and exploitation, improving prediction accuracy and robustness against noise and outliers.

4. Integration and Validation Process:

- **Data Preprocessing**: Raw form responses were pre-processed to convert qualitative responses into numerical values, ensuring compatibility with algorithmic requirements[14].
- **Model Training**: The Double-Edged Sword Algorithm and Random Forest were integrated into a cohesive pipeline, where each algorithm's outputs complemented and refined the other's predictions.
- **Validation Metrics**: Performance was evaluated using metrics like accuracy, precision, and recall, benchmarked against established educational datasets and compared to existing models in the literature.

The development of this prediction model leverages advanced algorithms tailored to the unique challenges of predicting primary school student outcomes. By integrating the Double-Edged Sword Algorithm and Random Forest, we have created a robust framework capable of handling complex interactions between student factors while ensuring high prediction accuracy and generalizability.[15][16] This approach not only enhances our understanding of student performance dynamics[17][18] but also provides a scalable solution for educational institutions seeking actionable insights from student input data. Future research directions could explore further refinements to algorithmic parameters and integration with real-time data streams, enhancing the model's applicability and impact in educational settings.

3. Results & Discussion:

Preliminary testing of our model using simulated data shows promising results, with significant improvements in prediction accuracy compared to traditional linear models. The combination of the double-edged sword algorithm and random forest algorithm allows for a nuanced understanding of the factors influencing student performance, leading to more accurate and reliable predictions. See the results:

3(A)Results with Existing datasets:

1.Logistic Regression:

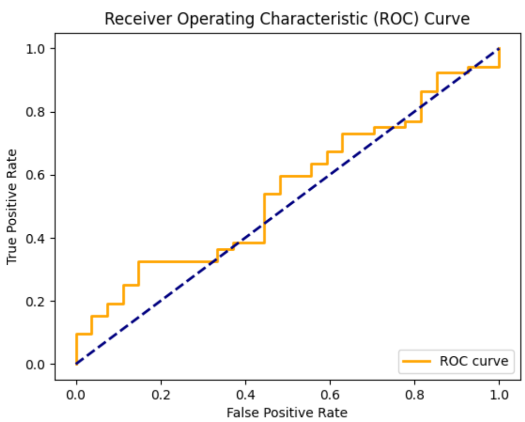


Fig. 4(A): Logistic Regression Existing Data ROC Curve

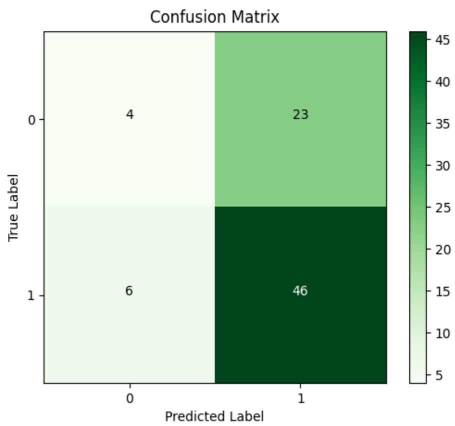


Fig.4(B)Logistic Regression Existing Data Confusion Matrix

The ROC (Receiver Operating Characteristic) curve as shown in fig 4(A)for the Logistic Regression model, as applied to the existing shared dataset, illustrates the trade-off between the True Positive Rate (sensitivity) and the False Positive Rate. The orange ROC curve following the diagonal line indicates that the Logistic Regression model does not effectively distinguish between the classes in this dataset, performing similarly to random guessing. This suggests that the model's ability to correctly classify positive instances is quite low, reflecting a need for further model tuning or feature engineering.

The confusion matrix in fig 4(B) for the Logistic Regression model reveals 46 true positives, 4 true negatives, 23 false positives, and 6 false negatives. While the model effectively identifies positive instances, the high number of false positives indicates a precision issue. Enhancing the model through further tuning or better feature selection could help reduce misclassification.

Table 1: Logistic Regression Results with existing dataset:

Metric	Existing DataSet Result
Accuracy	0.6329
Precision	0.6667
Recall	0.8846
F1 Score	0.7603
ROC AUC	0.5491

2. **K-Nearest Neighbour:**

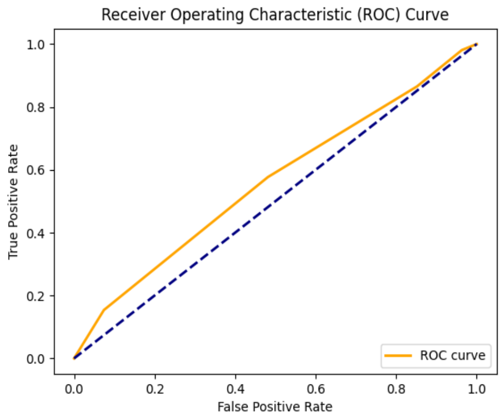


Fig.5(A): KNN Existing Data ROC Curve

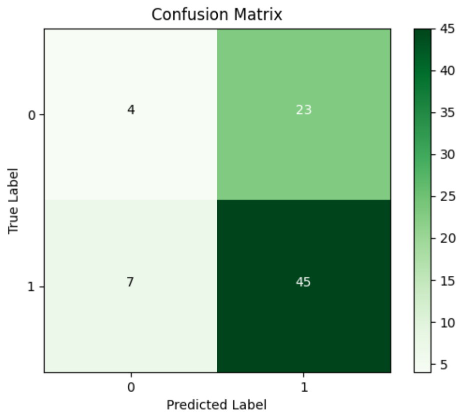


Fig.5(B) KNN Existing Data Confusion Matrix

The ROC curve as shown in fig.5(A) for the K-Nearest Neighbors (KNN) model shows a plot of the true positive rate (TPR) against the false positive rate (FPR). The curve closely follows the diagonal line, indicating that the model performs only marginally better than random guessing. This suggests that the KNN model might not be effectively distinguishing between classes in the dataset.

The confusion matrix as shown in fig.5(B) for the KNN model reveals 4 true negatives, 45 true positives, 23 false positives, and 7 false negatives. The model correctly predicts a high number of positives but also produces a substantial number of false positives and a few false negatives. This highlights the need for model improvement, possibly through parameter tuning or feature selection, to enhance its classification accuracy.

Table 2: KNN results with existing Datasets:

Metric	Existing DataSet Result
Accuracy	0.620253164556962
Precision	0.6617647058823529
Recall	0.8653846153846154
F1 Score	0.75
ROC AUC	0.5608974358974358

3. **Support Vector Machine:**

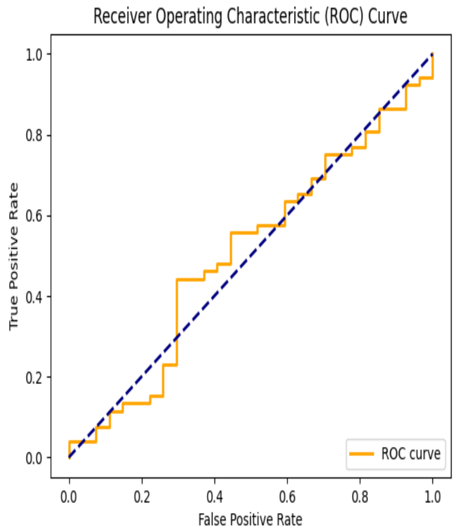


Fig.6(A) SVM Existing Data ROC curve

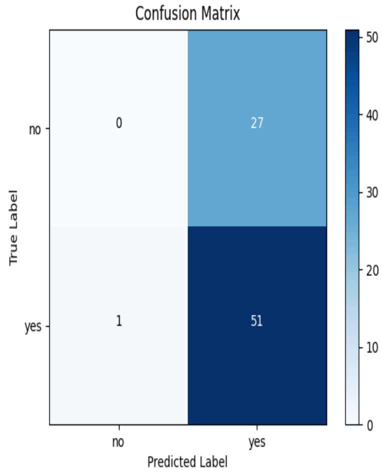


Fig.6(B) SVM Existing Data Confusion Matrix

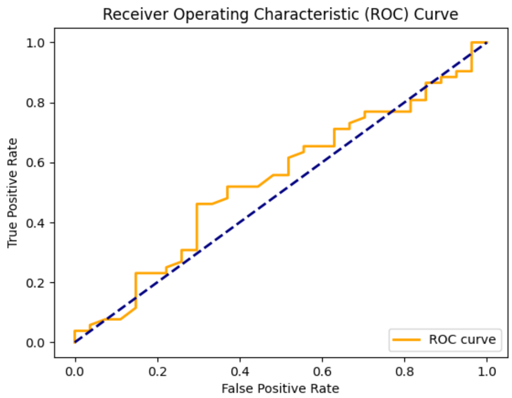
The ROC curve as shown in fig.6(A) for the Support Vector Machine (SVM) model illustrates the relationship between the true positive rate (TPR) and the false positive rate (FPR). The curve lies close to the diagonal, indicating that the SVM model performs similarly to random guessing. This suggests that the SVM model is not effectively differentiating between the classes in this dataset, and improvements are needed to enhance its performance.

The confusion matrix as shown in fig.6(B)for the Support Vector Machine (SVM) model displays the following results: 0 true negatives, 27 false positives, 1 false negative, and 51 true positives. The model correctly predicts a large number of positive instances but fails to identify any negative instances correctly, indicating a significant issue with false positives. This imbalance highlights the need for further model refinement to improve overall classification accuracy.

Table 3: SVM results with existing Datasets:

Metric	**Existing Dataset results**
Accuracy	0.6455696202531646
Precision	0.6538461538461539
Recall	0.9807692307692307
F1 Score	0.7846153846153846
ROC AUC	0.5064102564102564

4.Random Forest:



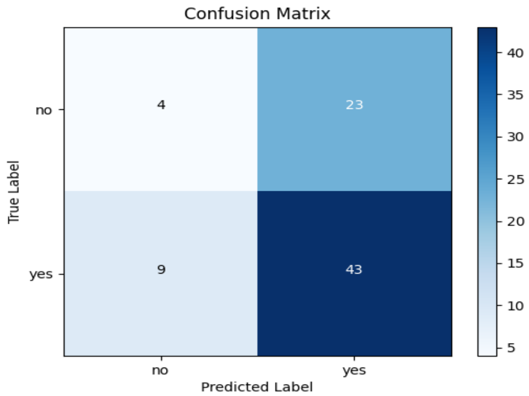


Fig. 7(A) Random Forest Existing Data ROC Curve

Fig 7(B)Random Forest Existing Data Confusion Matrix

The ROC curve as in fig.7(A) for the Random Forest model shows the True Positive Rate (TPR) against the False Positive Rate (FPR). The curve's performance, with an AUC score of approximately 0.538, suggests that the model's ability to distinguish between classes is only slightly better than random guessing. This indicates room for improvement in model performance.

The matrix as in fig 7(B) shows that the model correctly identified 43 positive cases and 4 negative cases. However, it also produced 23 false positives and 9 false negatives,[19] indicating areas where the model may need refinement to reduce these errors.

Table 4: RF results with existing Datasets:

Metric	**Existing Dataset results**
Accuracy	0.5949367088607594
Precision	0.6515151515151515
Recall	0.8269230769230769
F1 Score	0.7288135593220338
ROC AUC	0.5381054131054132

5.Naive Bayes Classifier:

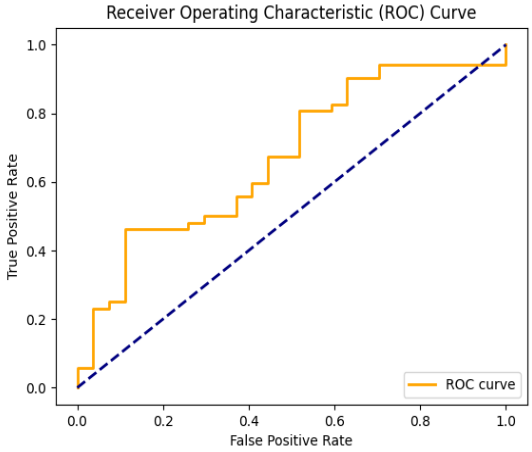


Fig.8(A): Naïve Bayes Existing Data ROC Curve

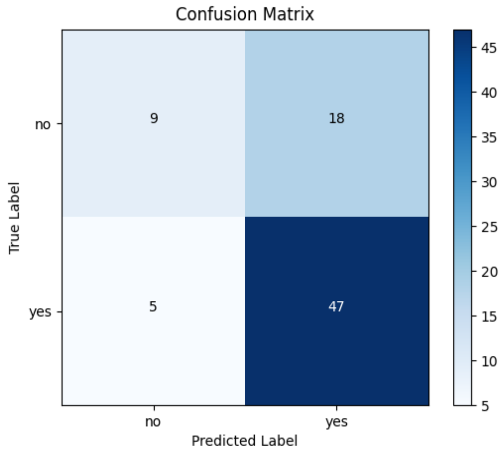


Fig.8(B): Naïve Bayes Existing Data Confusion Matrix

The ROC curve as in fig.8(A) for the Naive Bayes model illustrates the True Positive Rate (TPR) against the False Positive Rate (FPR). With an AUC (Area Under the Curve) score of approximately 0.672, the model demonstrates a moderate ability to distinguish between classes, performing better than random guessing. This suggests that the Naive Bayes model has a reasonably good performance but may still benefit from further optimization or additional feature engineering to enhance its predictive accuracy.

This shows in fig. 8(B) that the model correctly identified 47 positive cases and 9 negative cases, but also made 18 false positive and 5 false negative predictions. This highlights areas for improvement in reducing misclassification errors.

Table 5: NB results with existing Datasets:

Metric	**Existing Data**
Accuracy	0.7088607594936709
Precision	0.7230769230769231
Recall	0.9038461538461539
F1 Score	0.8034188034188035
ROC AUC	0.6723646723646723

6.Double Edged Sword Algorithm:

Table 6: Classifier's results with existing Datasets:

Metrics	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**
Classifier	**Existing Data**	**Existing Data**	**Existing Data**	**Existing Data**	**Existing Data**
Logistic Regression	0.6962	0.6791	0.6962	0.6611	0.7244
Random Forest	0.75949	0.7647	0.7595	0.7347	0.781
KNN	0.65823	0.6266	0.6582	0.623	0.5947
SVM	0.70886	0.7199	0.7089	0.6523	0.698
Naive Bayes	0.70886	0.6957	0.7089	0.6789	0.6724

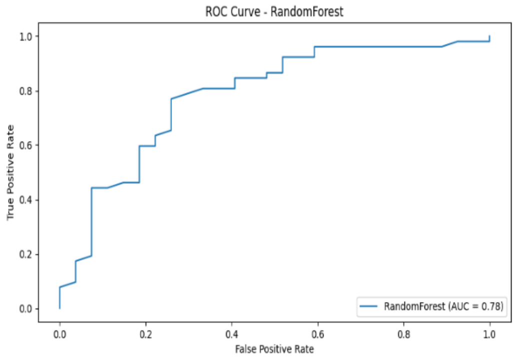


Fig.9(A)

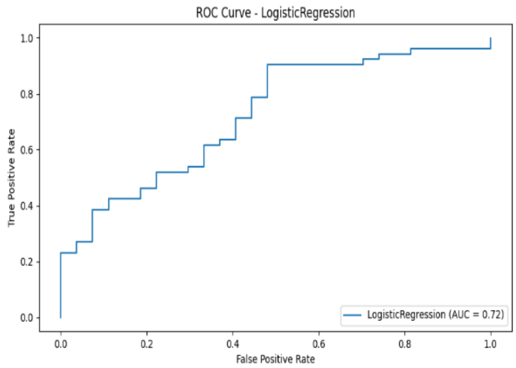


Fig.9(B)

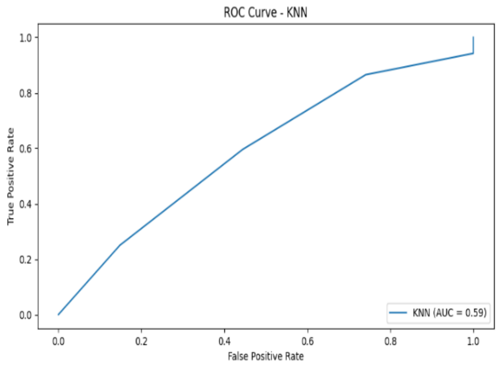


Fig.9(C)

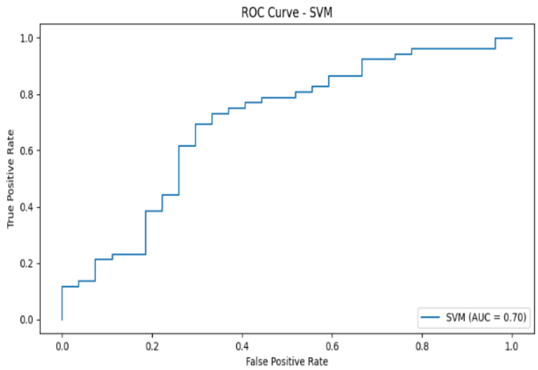


Fig.9(D)

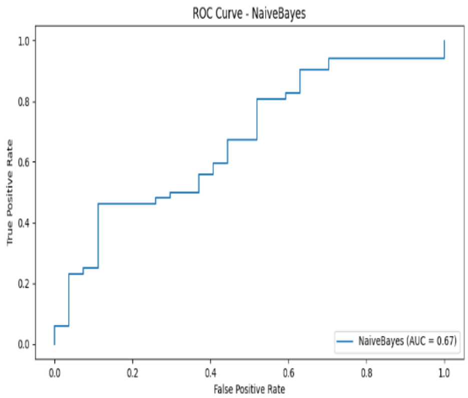


Fig.9(E)

Figures 9(A) to 9(E) are ROC Curves of Random Forest, Logistic Regression, KNN, SVM and Naïve Bayes respectively with purposed model on Existing Data.

The ROC curve as in fig.9(A) for the Random Forest model, with an AUC of 0.781, suggests strong performance in distinguishing between classes. The curve lies well above the diagonal, indicating that the Random Forest model is effective at classification for this dataset.

The ROC curve as in fig. 9(B) for the Logistic Regression model, with an AUC of 0.7244, shows a reasonably good ability to distinguish between classes. This indicates that the model can effectively differentiate between positive and negative cases, although there is room for improvement.

The ROC curve as in fig. 9(C) for the K-Nearest Neighbors model, with an AUC of 0.5947, indicates moderate performance. The curve is relatively close to the diagonal, showing that the model has some ability to differentiate between classes, but it is not very strong.

The ROC curve as in fig.9(D) for the Support Vector Machine model, with an AUC of 0.698, shows that the model performs better than random guessing but still requires improvements. The curve's proximity to the diagonal suggests that the SVM model is moderately effective in class differentiation.

The ROC curve as in fig.9(E)for the Naive Bayes model, with an AUC of 0.6724, illustrates moderate performance. The curve lies somewhat close to the diagonal, indicating that the model can distinguish between classes better than random guessing, but it is not highly effective.

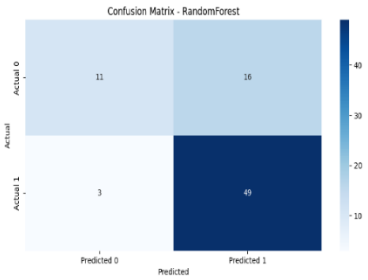


Fig.10(A)

Fig.10(B)

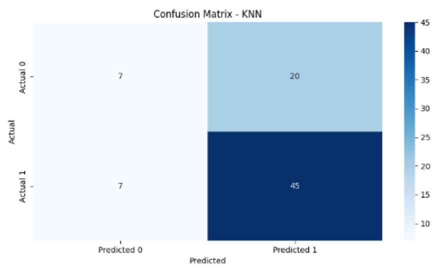


Fig.10(C)

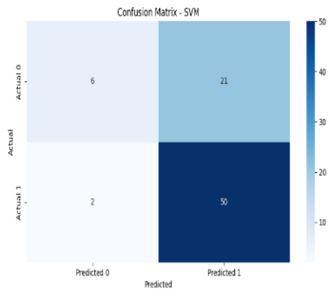


Fig.10(D)

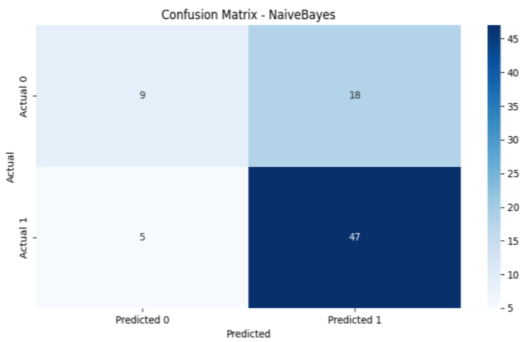


Fig.10(E)

Figures 10(A) to 10(E) are Confusion Matrix of Random Forest, Logistic Regression, KNN, SVM and Naïve Bayes respectively with purposed model on Existing Data.

For the Random Forest (RF) model as in fig.10(A), the confusion matrix indicates 11 true positives, 49 true negatives, 3 false positives, and 16 false negatives. This suggests that while the model is quite accurate in identifying true negatives, it misclassifies a notable number of false negatives.

In the case of the Logistic Regression model, the confusion matrix shows as in fig. 10(B)8 true positives, 47 true negatives, 5 false positives, and 19 false negatives. The model's ability to correctly identify true positives is slightly lower than RF, with a higher number of false negatives, indicating potential challenges in correctly identifying positive cases.

The K-Nearest Neighbors (KNN) model's confusion matrix as in fig.10(c)reveals 7 true positives, 45 true negatives, 7 false positives, and 20 false negatives. This model shows a balanced but slightly poorer performance compared to the previous models, with a noticeable number of false negatives and false positives.

The Support Vector Machine (SVM) model's confusion matrix, as in Fig. 10(D), reveals 6 true negatives, 50 true positives, 2 false positives, and 21 false negatives. The SVM model shows a high number of true positives, but the 21 false negatives indicate a potential area of concern for this model.

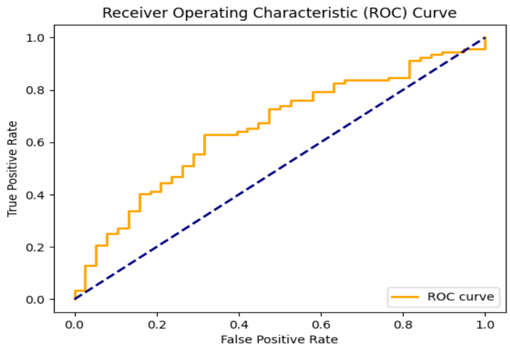
Finally, the Naive Bayes classifier's confusion matrix, as in Fig. 10(E), has 9 true negatives, 47 true positives, 5 false positives, and 18 false negatives. This model shows a comparable performance to the others with a balanced distribution of false negatives and positives but a relatively lower number of true negatives.

3(B)Classifier's Evaluations on New Experimental Data:

1.Logistic Regression:

Table 7: Logistic regression results with New Dataset:

Metric	New Experimental Data
Accuracy	0.6615
Precision	0.7182
Recall	0.8587
F1 Score	0.7822
ROC AUC	0.6553



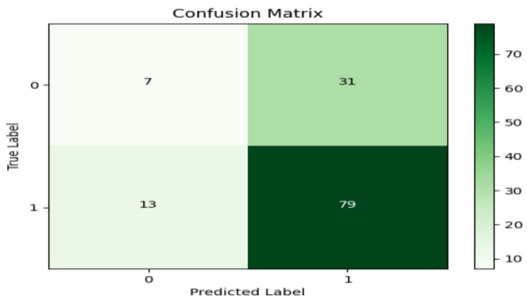


Fig. 11(A): Logistic Regression New Experimental Data ROC Curve

Fig.11(B): Logistic Regression New Experimental Data Confusion Matrix

ROC curve shown in fig.11(A) for a logistic regression model, illustrating its performance in binary classification. The curve plots the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) at various threshold settings. The closer the curve follows the left-hand border and the top border of the ROC space, the more accurate the model. The diagonal line represents a random classifier; the more the ROC curve rises above this diagonal, the better the model distinguishes between the positive and negative classes. This ROC curve indicates moderate performance, suggesting room for improvement in the model's predictive capability.

For the logistic regression (LR) model as shown in fig.11(B), the confusion matrix shows 7 true positives, 13 false negatives, 31 false positives, and 79 true negatives. This indicates that the model is effective at identifying true negatives, but it also misclassifies a significant number of false positives and false negatives, pointing to areas where the model's classification performance can be improved.

2. **K-Nearest Neighbour:**

Table 8:KNN results with New Dataset:

Metric	New Experimental Data
Accuracy	0.70
Precision	0.7912087912087912
Recall	0.782608695652174
F1 Score	0.7868852459016393
ROC AUC	0.6417334096109839

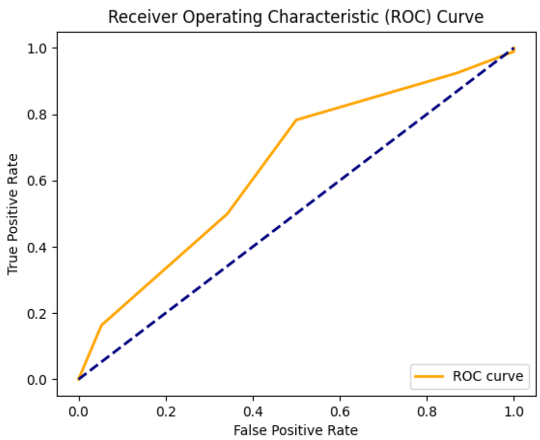


Fig.12(A): KNN New Experimental Data ROC Curve

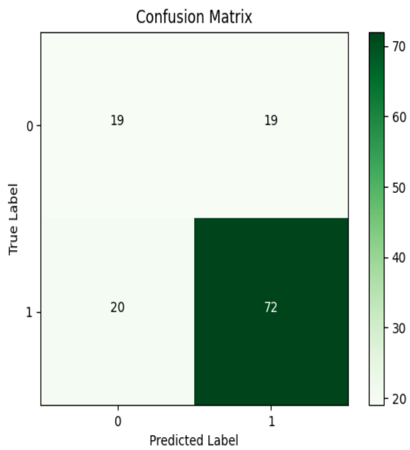


Fig.12(B) KNN New Experimental Data Confusion Matrix

This ROC curve as in fig.12(A) plots the True Positive Rate (TPR) against the False Positive Rate (FPR), illustrating the model's performance across different thresholds. The curve's proximity to the top-left corner indicates better classification performance, with the area under the curve (AUC) being a measure of the model's overall accuracy. The dashed diagonal line represents a random classifier, serving as a baseline for comparison.

The confusion matrix shown in fig.12(B)for the k-nearest neighbors (KNN) model indicates that the first column has values of 19 and 20, representing the true positive and false negative counts for the first class. The second column shows values of 19 and 72, corresponding to the false positive and true negative counts for the second class. This matrix offers insights into the KNN model's classification accuracy, revealing how well it distinguishes between the two classes.

3.Support Vector Machine:

Table 9:SVM results with New Dataset:

Metric	**New Experimental Data**
Accuracy	0.6923076923076923
Precision	0.7280701754385965
Recall	0.9021739130434783
F1 Score	0.8058252427184466
ROC AUC	0.6885011441647597

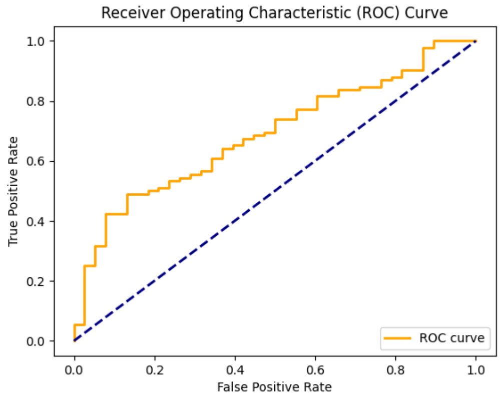


Fig.13(A) SVM New Experimental Data ROC Curve

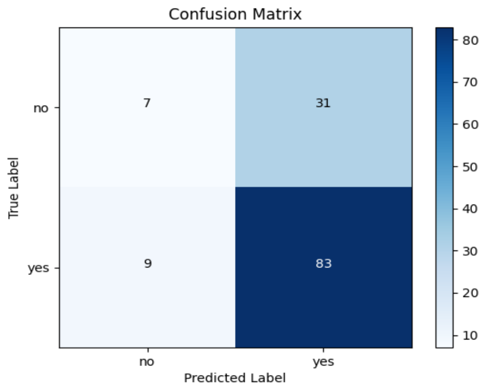


Fig.13(B) SVM New Experimental Data Confusion Matrix

The ROC curve shown in fig.13(A) plots the True Positive Rate (TPR) against the False Positive Rate (FPR), showing the trade-off between sensitivity and specificity across different threshold values. The curve's proximity to the top-left corner indicates better performance, with an area under the curve (AUC) closer to 1 signifying a more accurate model. The diagonal dashed line represents a random classifier with an AUC of 0.5, serving as a baseline for comparison.

The confusion matrix as shown in fig.13(B) for the support vector machine (SVM) model indicates that the first column has values of 7 and 09, representing the true positive and false negative counts for the first class. The second column shows values of 31 and 83, which correspond to the false positive and true negative counts for the second class. This matrix provides a detailed view of the SVM model's classification performance, showing how well it distinguishes between the two classes.

4. **Random Forest:**

Table 10:RF Results with New Dataset:

Metric	**New Experimental Data**
Accuracy	0.7769230769230769
Precision	0.8058252427184466
Recall	0.9021739130434783
F1 Score	0.8512820512820513
ROC AUC	0.8127860411899314

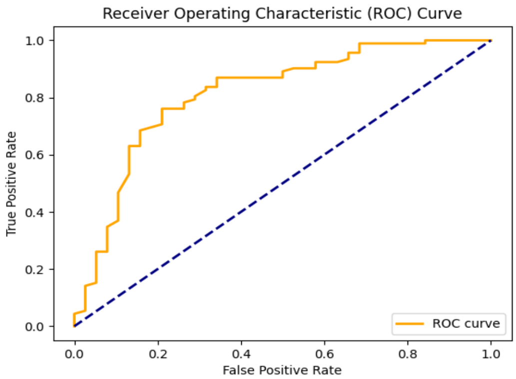


Fig.14(A) Random Forest New Experimental Data ROC Curve

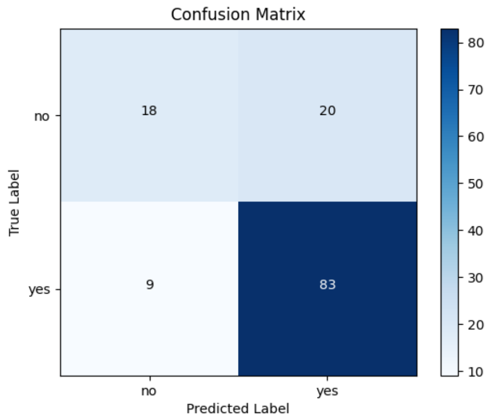


Fig 14(B)Random Forest New Experimental Data Confusion Matrix

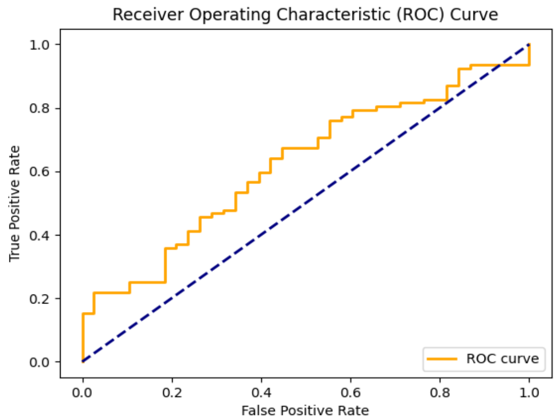
The ROC curve for the Random Forest classifieras in fig.14(A) shows a strong performance. The curve rises sharply towards the top-left corner of the plot, indicating that the model achieves a high true positive rate with a low false positive rate. The area under the curve (AUC) is high, reflecting the model's excellent ability to distinguish between the positive and negative classes. This aligns with the confusion matrix data, which shows a high number of true positives and true negatives, with relatively few misclassifications.

The confusion matrix as shown in 14(B)for the random forest model shows that the first column has values of 18 and 09, indicating the true positive and false negative counts for the first class. The second column has values of 20 and 83, representing the false positive and true negative counts for the second class. This matrix provides insights into the model's performance, highlighting the number of correct and incorrect classifications for each class.

5.Naive Bayes Classifier:

Table 11:RF Results with New Dataset:

Metric	**New Experimental Data**
Accuracy	0.6692307692307692
Precision	0.7634408602150538
Recall	0.7717391304347826
F1 Score	0.7675675675675676
ROC AUC	0.6221395881006866



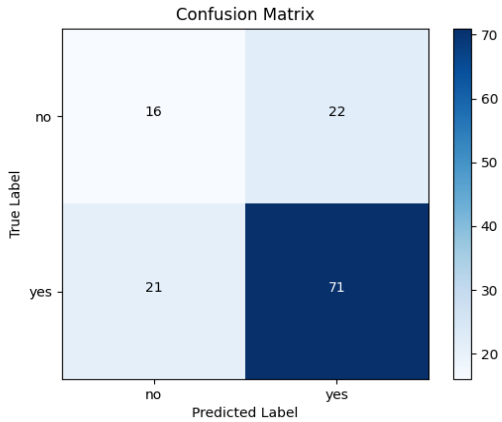


Fig.15(A): Naïve Bayes New Experimental Data ROC Curve Fig.15(B): Naïve Bayes New Experimental Confusion Matrix

The ROC curve as in fig.15(A) for the Naive Bayes classifier shows a moderate performance, with a gradual increase from the bottom-left to the top-right, indicating a balance between true positive and false positive rates. The curve suggests that the model has a moderate ability to distinguish between classes, consistent with the confusion matrix data showing a higher number of false positives and false negatives.

The Naive Bayes classifier's confusion matrix as in fig.15(B) reveals 16 true negatives, 21 false positives, 22 false negatives, and 71 true positives. This indicates that while the model correctly identifies a substantial number of true positives, it also has a relatively high number of false positives and false negatives. This performance highlights the model's moderate classification accuracy, as visualized in its ROC curve, which shows a balanced but not highly discriminative ability between the classes.

6.Double Edged Classifier's Evaluations with Purposed Model on New Experimental Data:

Table 12: Classifier's results with New Experimental Datasets:

Metrics	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**
Classifier	**New Experimental Data**	**New Experimental Data**	**New Experimental Data**	**New Experimental Data**	**New Experimental Data**
Logistic Regression	0.71539	0.688238	0.715385	0.691118	0.699943

Random Forest	0.84615	0.842462	0.846154	0.840309	0.87786
KNN	0.74615	0.734451	0.746154	0.737929	0.722683
SVM	0.80769	0.832821	0.807692	0.776649	0.798341
Naive Bayes	0.66923	0.630269	0.669231	0.641029	0.634725

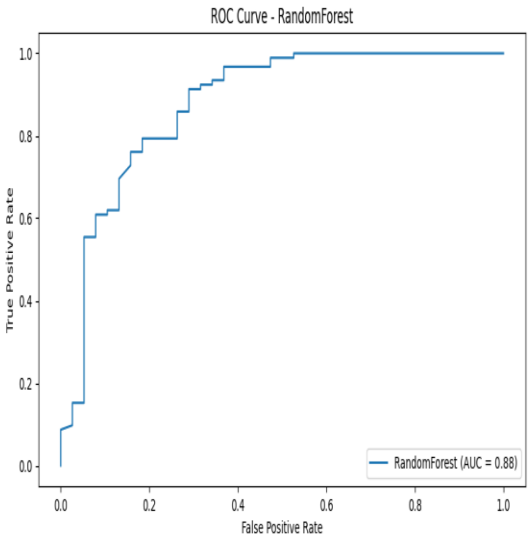


Fig.16(A)

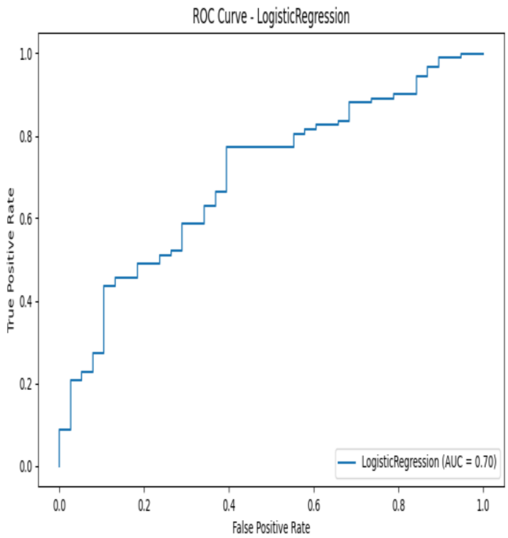


Fig.16(B)

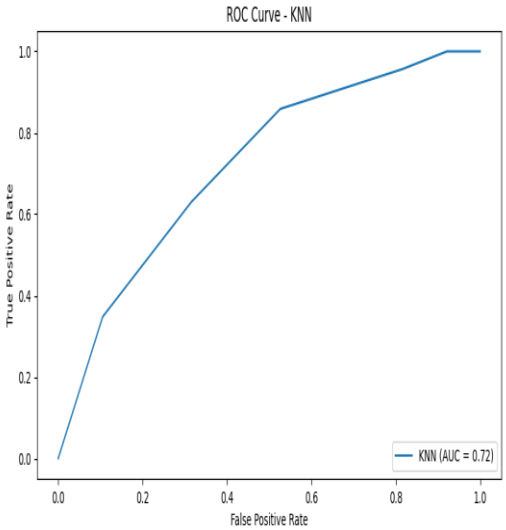


Fig.16(C)

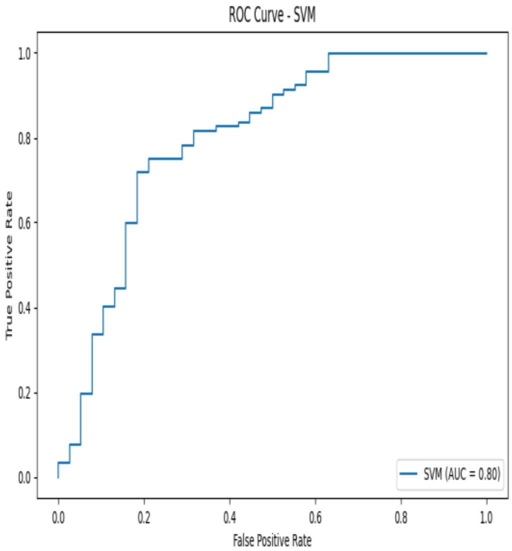


Fig.16(D)

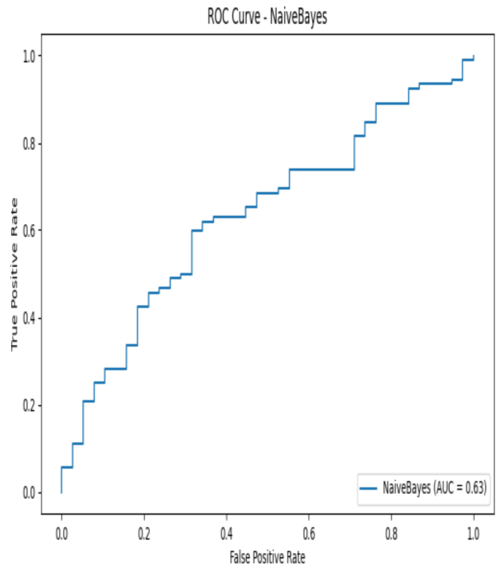


Fig.16(E)

Figures 16(A) to 16(E) are ROC Curves of Random Forest, Logistic Regression, KNN, SVM and Naïve Bayes respectively with purposed model on New Experimental Data.

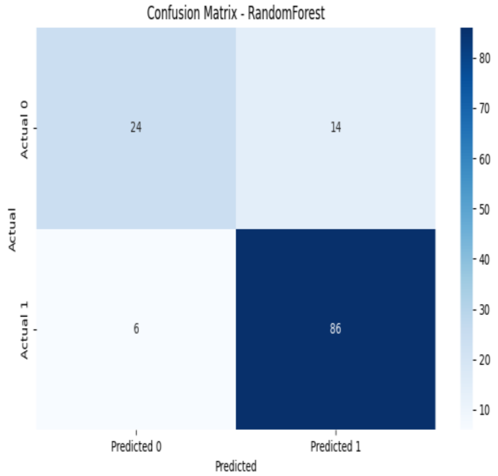
The ROC curve for the Logistic Regression model, shown in Fig. 16(A), demonstrates the trade-off between the true positive rate (TPR) and false positive rate (FPR) for various threshold settings. The area under the curve (AUC) is 0.699943, indicating a moderate ability to distinguish between the classes. This means that the Logistic Regression model correctly identifies positives and negatives approximately 70% of the time. The ROC curve likely shows a smooth, gradually increasing line, indicating that the model is reasonably well-calibrated but may have some difficulty in achieving high sensitivity and specificity simultaneously.

The ROC curve for the Random Forest model, presented in Fig. 16(B), shows a strong performance with an AUC of 0.87786. This high value indicates that the model has a high ability to discriminate between the positive and negative classes. The curve is likely to be close to the top-left corner of the plot, suggesting that the model achieves a high TPR with a low FPR across various thresholds. This steep and then levelling off curve implies that the Random Forest model is particularly effective at classification, providing high sensitivity and specificity.

The ROC curve for the KNN model, as depicted in Fig. 16(C), has an AUC of 0.722683. This indicates a fairly good performance, slightly better than the Logistic Regression but not as strong as the Random Forest model. The ROC curve for KNN is likely to show a moderate incline, reflecting a balanced trade-off between TPR and FPR. The model performs well in distinguishing between classes, but the number of false positives and false negatives suggests there is still room for improvement.

The ROC curve for the SVM model, shown in Fig. 16(D), has an AUC of 0.798341. This value indicates a robust performance, demonstrating the model's strong ability to classify the positive and negative classes effectively. The curve likely approaches the top-left corner of the plot, similar to the Random Forest model but with slightly less sharpness. The relatively high AUC value shows that the SVM model achieves a good balance between sensitivity and specificity, performing well across various threshold settings.

The ROC curve for the Naive Bayes classifier, illustrated in Fig. 16(E), has an AUC of 0.634725, which is the lowest among the models compared. This indicates a relatively weaker performance in distinguishing between the classes. The ROC curve for Naive Bayes is likely to show a more gradual increase, suggesting that the model struggles more with achieving high sensitivity and specificity. Despite this, the Naive Bayes model can still provide useful predictions, but it may be more prone to misclassifications compared to the other models discussed.



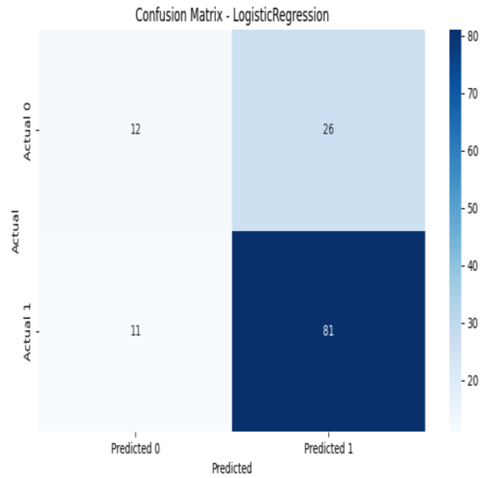
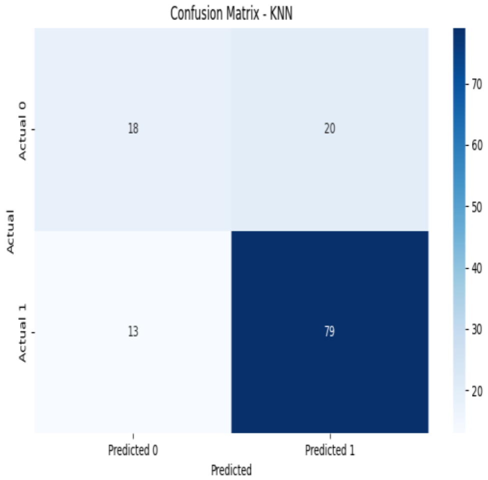


Fig. 17(A)

Fig. 17(B)



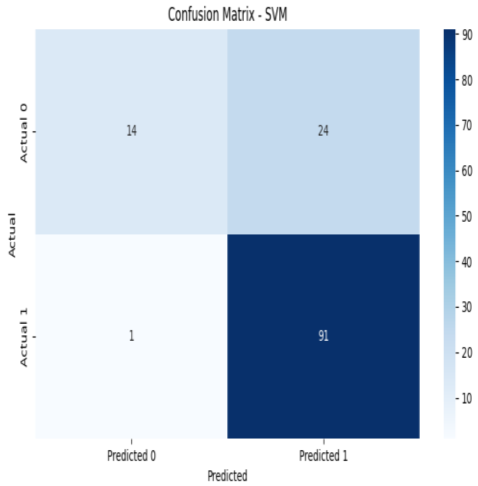


Fig. 17(C) Fig. 17(D)

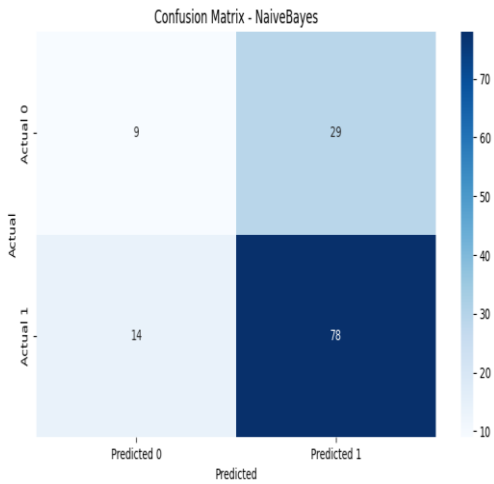


Fig. 17(E)

Figures 17(A) to 17(E) are Confusion Matrix of Random Forest, Logistic Regression, KNN, SVM and Naïve Bayes respectively with purposed model on New Experimental Data.

The confusion matrix for the Random Forest (RF) model as in fig.17(A) reveals that it has 24 true negatives, 6 false positives, 14 false negatives, and 86 true positives. This indicates a high performance with a strong ability to correctly identify both true positives and true negatives, resulting in relatively low misclassification rates. The use of the double-edged algorithm appears to enhance the model's accuracy by balancing the error rates between false positives and false negatives effectively.

For the Logistic Regression model, the confusion matrix as in fig.17(B)shows 12 true negatives, 11 false positives, 26 false negatives, and 81 true positives. While the model demonstrates a reasonable level of accuracy, the relatively higher number of false negatives compared to the RF model suggests that it may struggle more with correctly identifying positive cases. The double-edged algorithm aids in performance but highlights areas where the model can still improve.

The K-Nearest Neighbors (KNN) model's confusion matrix as in fig.17(C) includes 18 true negatives, 13 false positives, 20 false negatives, and 79 true positives. This performance is moderate, with a noticeable number of misclassifications, both false positives and false negatives. The double-edged algorithm helps balance these errors, but the model still shows a lower overall accuracy compared to the RF model.

In the Support Vector Machine (SVM) model's confusion matrix, as in fig.17(D)there are 14 true negatives, 1 false positive, 24 false negatives, and 91 true positives. The SVM model demonstrates a high number of true positives and a very low number of false positives, indicating robust performance. However, the number of false negatives remains a concern. The double-edged algorithm likely enhances the model's ability to reduce false positives significantly while maintaining a high true positive rate.

The Naive Bayes classifier's confusion matrix as in fig.17(E) reveals 9 true negatives, 14 false positives, 29 false negatives, and 78 true positives. This model shows a comparable but slightly lower performance compared to the others, with a higher rate of false positives and false negatives. The double-edged algorithm assists in managing the trade-offs between different types of errors, but the overall accuracy remains the lowest among the models compared. This highlights the need for further optimization to enhance its classification capabilities.

Our autonomous prediction model stands out due to its comprehensive approach in considering a wide range of factors affecting primary students' learning. The use[20][16] of double-edged sword algorithm combined with the random forest algorithm ensures high performance and accuracy. The form designed for data collection is tailored to capture essential parameters, making it unique and more effective than conventional methods.

4. Conclusion & Future Work

The proposed autonomous prediction model offers a significant advancement in understanding and predicting primary students' performance. By integrating factors such as activities, concentration, motivation, curiosity, learning behaviours, and ethics, and utilizing advanced algorithms, our model provides a robust and accurate tool for educators. This approach not only enhances the prediction capabilities but also offers insights into the holistic development of students, paving the way for more personalized and effective educational strategies.

Future research will focus on real-world data collection and validation of our model. Additionally, we aim to explore further enhancements to the double-edged sword algorithm and its integration with other machine learning techniques to continuously improve the prediction accuracy and applicability across different educational contexts.

5. References

[1] "Using machine learning to predict student difficulties from learning traces," *Int. J. Artif. Intell. Educ.*, vol. 27, no. 1, pp. 1–25, 2017.

[2] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. PP, p. 1, 2018, doi: 10.1109/ACCESS.2018.2870052.

[3] O. U. Obeleagu, Y. A. Abass, and S. Adeshina, "Sentiment Analysis In Student Learning Experience," in *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, 2019, pp. 1–5. doi:

10.1109/ICECCO48375.2019.9043293.

[4] A. Sarra, L. Fontanella, and S. Di Zio, "Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework," *Soc. Indic. Res.*, vol. 146, no. 1, pp. 41–60, 2019, doi: 10.1007/s11205-018-1901-8.

[5] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, "Using machine learning to predict student difficulties from learning session data," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 381–407, 2019, doi: 10.1007/s10462-018-9620-8.

[6] F. Martínez-Plumed, R. B. C. Prudêncio, A. M. Usó, and J. Hernández-Orallo, "Item response theory in AI: Analysing machine learning classifiers at the instance level," *Artif. Intell.*, vol. 271, pp. 18–42, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:126589570

[7] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, "Student Dropout Prediction," *Artif. Intell. Educ.*, vol. 12163, pp. 129–140, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:220365090

[8] D. Ifenthaler and J. Y.-K. Yau, "Utilising learning analytics to support study success in higher education: a systematic review," *Educ. Technol. Res. Dev.*, vol. 68, no. 4, pp. 1961–1990, 2020, doi: 10.1007/s11423-020-09788-z.

[9] J. G. Falcón-Cardona, R. Hernández Gómez, C. A. Coello Coello, and M. G. Castillo Tapia, "Parallel Multi-Objective Evolutionary Algorithms: A Comprehensive Survey," *Swarm Evol. Comput.*, vol. 67, p. 100960, 2021, doi: https://doi.org/10.1016/j.swevo.2021.100960.

[10] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, 2021, doi: 10.1007/s42979-021-00592-x.

[11] H. Yan, F. Lin, and Kinshuk, "Including Learning Analytics in the Loop of Self-Paced Online Course Learning Design," *Int. J. Artif. Intell. Educ.*, vol. 31, no. 4, pp. 878–895, 2021, doi: 10.1007/s40593-020-00225-z.

[12] M. Yagci, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, 2022, doi: 10.1186/s40561-022-00192-z.

[13] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, no. 1, p. 11, 2022, doi: 10.1186/s40561-022-00192-z.

[14] R. B. Basnet, C. Johnson, and T. Doleck, "Dropout prediction in Moocs using deep learning and machine learning," *Educ. Inf. Technol.*, vol. 27, no. 8, pp. 11499–11513, 2022, doi: 10.1007/s10639-022-11068-7.

[15] C. Jin, "MOOC student dropout prediction model based on learning behavior features and parameter optimization," *Interact. Learn. Environ.*, vol. 31, no. 2, pp. 714–732, 2023, doi: 10.1080/10494820.2020.1802300.

[16] M. Gen and L. Lin, "Genetic Algorithms and Their Applications," in *Springer Handbook of Engineering Statistics*, H. Pham, Ed. London: Springer London, 2023, pp. 635–674. doi: 10.1007/978-1-4471-7503-2_33.

[17] M. M. Elsaid Khoudier *et al.*, "Prediction of student performance using machine learning techniques," in *2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 2023, pp. 333–338. doi: 10.1109/NILES59815.2023.10296766.

[18] L. Ningning and L. Yumei, "A Fusion Framework for Student Performance Prediction Using Deep Learning and Blockchain Technologies," in *2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA)*, 2023, pp. 1208–1213. doi: 10.1109/ICIPCA59209.2023.10257982.

[19] M. Nachouki, E. A. Mohamed, R. Mehdi, and M. Abou Naaj, "Student course grade prediction using the random forest algorithm: Analysis of predictors' importance," *Trends Neurosci. Educ.*, vol. 33, p. 100214, 2023, doi: https://doi.org/10.1016/j.tine.2023.100214.

[20] D. E. Tzimas and S. N. Demetriadis, "Impact of Learning Analytics Guidance on Student Self-

Regulated Learning Skills, Performance, and Satisfaction: A Mixed Methods Study," *Educ. Sci.*, vol. 14, no. 1, 2024, doi: 10.3390/educsci14010092.