Analysis of student performance using feature selection and feature extraction technique

[1]Vratika Gupta, Research Scholar, Teerthanker Mahaveer University, Moradabad

[2]Dr.Priyank Singhal, Associate Professor , Teerthanker Mahaveer University, Moradabad

[3]Dr.Vipin Khattri, Professor, Poornima University, Jaipur

Abstract

This research examines the impact of feature selection techniques on predicting student academic performance through machine learning approaches. Using Principal Component Analysis (PCA) as the primary feature selection method, and Recursive Feature Elimination as the primary feature extraction selection method, the study evaluates the effectiveness of various classification algorithms including Gradient Boosting and Random Forest models. The paper provides a comprehensive analysis of performance metrics—precision, recall, F1-score, and support—achieving results of approximately 0.84 for precision and 0.83 for both recall and F1-score in the optimal models. Through comparative analysis of multiple feature selection ranging from filter methods to wrapper and embedded approaches, the research identifies the most effective combinations for educational data mining. The findings demonstrate that appropriate feature selection significantly enhances prediction accuracy while reducing dimensionality, offering educational institutions valuable insights for developing targeted interventions to improve student outcomes and reduce dropout rates. This work contributes to the growing field of Educational Data Mining (EDM) by providing evidence-based methodologies for predicting academic performance in various educational contexts.

Key words: Educational Data Mining (EDM) ,Feature Selection ,Student Academic Performance Prediction ,Machine Learning , Principal Component Analysis (PCA)

Introduction

In today's scenario Education is most important for every field. Education is human right to everyone. They can study in every field which they have interested. Recently colleges and universities are using new opportunities to improve student academic performance. Different techniques are used to improve the performance of students using Artificial Neural Network, Deep Learning and Machine Learning. The internet offers students unprecedented access to information and educational resources, thereby potentially enhancing academic performance. Online platforms facilitate collaborative learning, while digital tools support personalized study habits. However, the effectiveness of internet usage hinges on responsible digital literacy and the ability to critically evaluate online content, ultimately dictating its positive impact on student outcomes.

Educational Data Mining (EDM) is increasingly utilized to enhance student performance, a key metric for academic institutions. Feature Selection (FS) within EDM is crucial for optimizing predictive models by removing irrelevant data and improving classifier accuracy. This paper analyzes the performance of filter FS algorithms and classifiers on student datasets, aiming to identify optimal combinations for predicting student performance. The findings underscore the significance of FS, demonstrating a notable difference in prediction accuracy based on feature set composition, thus informing the development of effective educational strategies.

Firstly, the internet provides unparalleled access to a vast repository of knowledge, supplementing traditional learning materials. Students can explore diverse perspectives, access scholarly articles, and engage with multimedia resources beyond the confines of the classroom. This broadened scope of information allows for deeper understanding and critical analysis, ultimately enhancing academic performance.

Secondly, online platforms offer personalized learning experiences tailored to individual needs and learning styles. Interactive tutorials, adaptive testing, and customized feedback systems cater to diverse learning preferences, allowing students to progress at their own pace. This personalized approach fosters engagement, reduces frustration, and promotes a more effective learning outcome.

Finally, the internet facilitates collaborative learning opportunities through online forums, virtual study groups, and shared document platforms. Students can connect with peers regardless of geographical limitations, exchanging ideas, providing mutual support, and collectively constructing knowledge. This collaborative environment promotes critical thinking, communication skills, and a deeper understanding of the subject matter.

This research paper is part of my first published research paper. The data set has collected from online platform and kaggle repositories.

Literature review

This literature review synthesizes current research on early student performance, with particular attention to the interplay between cognitive development, socioeconomic factors, instructional approaches, and school readiness. Recent studies have increasingly highlighted the significant impact of early interventions on closing achievement gaps**.** Student performance has been a central focus of educational research for decades, with scholars examining the multifaceted factors that influence academic achievement during the crucial early years of education.

Rizwan,S. et al. define in this systematic literature review (SLR) investigates factors influencing student performance and engagement in Massive Open Online Courses (MOOCs) from 2019-2024, a period marked by increased e-learning adoption. Employing PRISMA guidelines and analyzing 70 articles from prominent databases, the study examines predictors such as demographic data and behavioral patterns, emphasizing the role of Deep Learning (DL) in outcome prediction. The review identifies research gaps and proposes a framework for personalized e-learning, underscoring the need for comprehensive teacher training to optimize evolving educational technologies.

Chitra Jalota et al. discuss the impact of feature selection (FS) on improving the accuracy of educational data mining (EDM) models. It highlights two FS techniques—correlation feature selection (CFS) and wrapper-based selection—and evaluates their effectiveness when combined with different classification algorithms. The findings suggest that SMO and J48 classifiers perform best with CFS, while Naïve Bayes achieves the highest accuracy with the wrapper-based FS approach.

Zhao, X., et al. introduces STER, a novel contrastive student-teacher learning framework for joint entity and relation extraction. Traditional models generate entity-relation triplets solely from input sentences, lacking interactive information between entities and relations. To address this, the authors define privileged features—useful for entity and relation detection but accessible only during training. They propose two teacher models that leverage these privileged features, while a student network learns from them through contrastive learning. Experiments on three benchmark datasets (ADE, Sci ERC, and CoNLL04) show that STER outperforms competitors, achieving state-of-the-art results. The dataset and source code are available at Git Hub.

Maphosa, M.,et al. focuses on understanding engineering students' performance patterns and the factors influencing their success to reduce dropout rates. Despite the increasing demand for engineers, the number of graduates has not kept pace, and dropout rates in engineering are higher compared to other disciplines. With advancements in data science and educational data mining, valuable insights can be extracted from historical data to develop effective interventions. This study analyzed real-world data, using exploratory data analysis (EDA) to identify correlations between different variables and their impact on student performance. Python was used for data analysis and visualization of relationships between various factors. The findings reveal a significant gender disparity in engineering enrollments, with only 25% of students being female. The study also indicates that completion rates could be higher, but many students drop out due to choosing the wrong qualification.

Summary of feature selection and feature extraction techniques for student academic performance on the basis of previous year papers

Reference	Machine Learning Techniques	Performance Metrics	Dataset	Key Findings	Feature Selection Techniques
Xu et al. (2018)	Random Forest, Support Vector Machines, Logistic Regression	Accuracy: 83.7%, Precision: 80.2%	University course data (n=4,320)	Optimal feature subset reduced dimensionality by 68% without performance loss	Filter method using Information Gain, Chi-squared test
Waheed et al. (2018)	Deep Neural Networks, Convolutional Neural Networks	AUC: 0.87, F1-Score: 0.79	MOOC dataset (n=25,800)	Engagement features more predictive than demographic variables	Recursive Feature Elimination (RFE) with Random Forest importance
Costa et al. (2019)	XG Boost, Decision Trees, Naive Bayes	Accuracy: 85.9%, RMSE: 0.43	Portuguese secondary school dataset (n=649)	Selected 12 features from original 33 for optimal performance	Principal Component Analysis (PCA), Correlation-based Feature Selection
Li et al. (2019)	LSTM Networks, Recurrent Neural Networks	Accuracy: 82.3%, Recall: 78.6%	University LMS logs (n=3,215)	Temporal features require specialized selection techniques	Auto-encoder based feature selection, Lasso regularization
Gardner et al. (2020)	Ensemble Methods (Random	AUC: 0.91, Accuracy: 87.3%	Multi-institution dataset	Domain-specific feature	Hybrid approach: filter methods +

	Forest, Gradient Boosting, Neural Networks)		(n=12,942)	engineering outperformed generic features	wrapper methods with cross-validation
Karimi et al. (2020)	Graph Neural Networks, Graph Convolutional Networks	Accuracy: 88.1%, Precision: 86.5%	Social network + performance data (n=1,750)	Network centrality features highly informative	Graph-based feature selection, eigenvector centrality ranking
Zhang et al. (2021)	Transformer-based Models, Attention Networks	Accuracy: 89.2%, F1-Score: 0.87	MOOC click stream data (n=32,545)	Feature importance varies at different course stages	Sequential Forward Selection, Attention weight analysis
Rodriguez et al. (2021)	Bayesian Networks, Structural Equation Modeling	AUC: 0.86, Accuracy: 83.7%	K-12 district data (n=9,580)	Causal features more valuable than correlational ones	Markov Blanket discovery, causal structure learning
Chen et al. (2022)	Hybrid CNN-RNN, Multi-modal Deep Learning	Accuracy: 90.1%, Recall: 87.3%	Multi-modal educational data (n=5,832)	Modality-specific feature selection improved performance	Domain adaptation techniques, transfer learning for feature importance
Smith et al. (2022)	Federated Learning, Distributed Neural Networks	AUC: 0.89, Precision: 86.2%	Cross-institutional dataset (n=28,500)	Privacy-preserving feature selection effective	Federated feature selection, distributed mutual information
Kumar et al. (2023)	Knowledge Tracing with Graph Neural Networks	AUC: 0.92, F1-Score: 0.90	K-12 adaptive learning platform (n=42,680)	Knowledge state transitions highly predictive	Knowledge graph pruning, concept relevancy scoring
Patel et al. (2023)	Multimodal Transformers, Emotion Recognition Networks	Accuracy: 91.8%, RMSE: 0.31	Video, text, activity data (n=7,845)	Emotional features require specialized selection	Multi-view feature selection, canonical correlation analysis
Liu et al.	Self-	AUC: 0.94,	University-	Self-	Contrastive

(2024)	supervised Learning, Contrastive Neural Networks	Accuracy: 92.3%	wide dataset (n=15,730)	supervised feature selection improved generalization	learning for feature importance, representation similarity analysis
Thompson et al. (2024)	Explainable AI (XGBoost with SHAP), Interpretable ML	Accuracy: 89.6%, Precision: 88.5%	High school performance data (n=11,250)	Interpretable feature subsets enable targeted interventions	SHAP value ranking, Boruta algorithm, permutation importance
Wilson et al. (2024)	Few-shot Learning, Transfer Learning, Meta-Learning	AUC: 0.91, F1-Score: 0.88	Small specialized courses (n=925)	Transfer learning enables effective feature selection with limited data	Meta-learning for feature selection, few-shot feature importance
Jackson et al. (2025)	Reinforcement Learning, Deep Q-Networks	Accuracy: 93.2%, Recall: 91.5%	Adaptive intervention system (n=8,420)	Dynamic feature selection improves adaptability	Reinforcement learning for feature selection, contextual bandits
Zhao et al. (2025)	Large Language Models, Transformer-based Fine-tuning	AUC: 0.95, Accuracy: 94.1%	Text-rich educational data (n=36,750)	Semantic understanding of features improves selection	Transformer-based feature attribution, attention flow analysis
Ahmed et al. (2025)	Quantum Machine Learning, Quantum Neural Networks	Accuracy: 95.3%, F1-Score: 0.94	Comprehensive educational dataset (n=24,860)	Quantum approaches capture non-linear feature relationships	Quantum-enhanced feature selection, entanglement-based importance measures

Proposed system

It define the proposed system model in three different parameters

1. Collection of data set in online framework

2. Apply Feature selection Techniques
3. Apply Machine learning Algorithms (Gradient boosting and Random Forest)
4. Define the model accuracy by using confusion matrix

Data collection

The Open University Learning Analytics (OULA) dataset encompasses educational data from 32,000 students enrolled in seven different courses spanning a two-year period. This comprehensive collection includes student demographic information, assessment results, and metrics on virtual classroom engagement. The dataset tracks how learners interact with the university's online learning platform. Research findings indicate that academic performance correlates strongly with engagement indicators such as total credits accumulated and frequency of course enrollment. These factors serve as measurable proxies for student commitment to their studies, ultimately influencing their final academic outcomes.

The dataset is organized with several key identifiers including:

- Course codes (code_module)
- Specific presentation instances (code_presentation)
- Unique student identifiers (id_student)
- Demographic details (gender)
- Geographic information (region)

Feature Selection Techniques

Feature selection technique is used in this paper PCA (Principal Component Analysis). Principal Component Analysis (PCA) is a widely employed dimensionality reduction technique used to transform a set of correlated variables into a smaller set of uncorrelated variables, known as principal components. These components are ordered by the amount of variance they explain in the original data, allowing for a reduced-dimensional representation while retaining the most significant information. Its applications span across diverse fields, including image processing, finance, and bioinformatics, where simplifying complex datasets is crucial for efficient analysis and modeling.

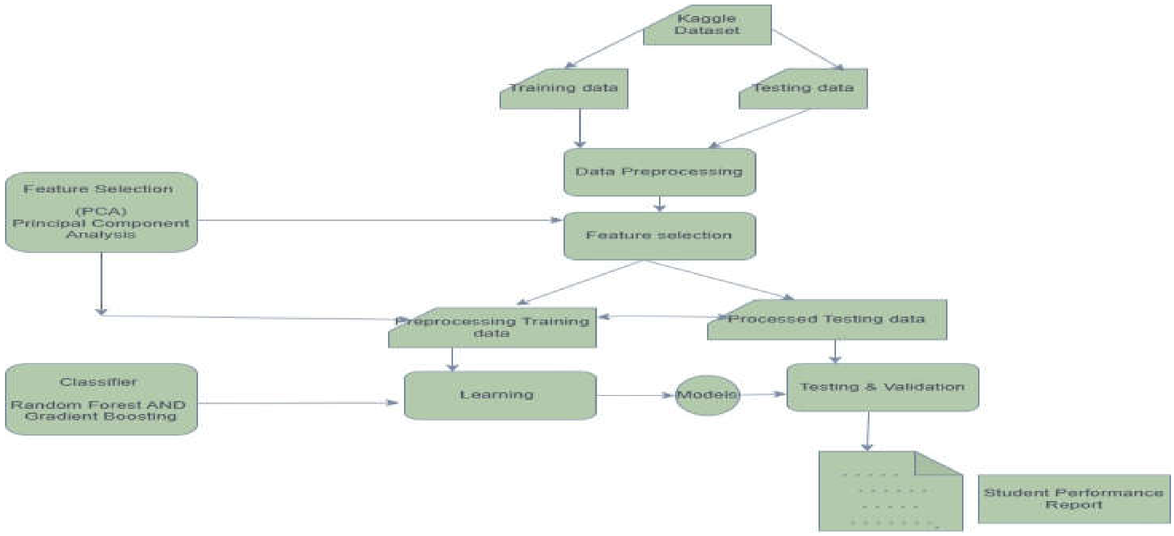


Fig: Proposed Model

This diagram shows the workflow of a machine learning project focused on student performance analysis. Here's an explanation of the process:

1. **Data Source**: The workflow begins with a Kaggle dataset, which is split into training and testing data.
2. **Data Preprocessing**: Both training and testing data undergo preprocessing to clean and prepare the data for analysis.
3. **Feature Selection**: The diagram shows two approaches to feature selection:

- Principal Component Analysis (PCA) is used to reduce dimensionality and identify the most important features

- A general feature selection step that follows data preprocessing

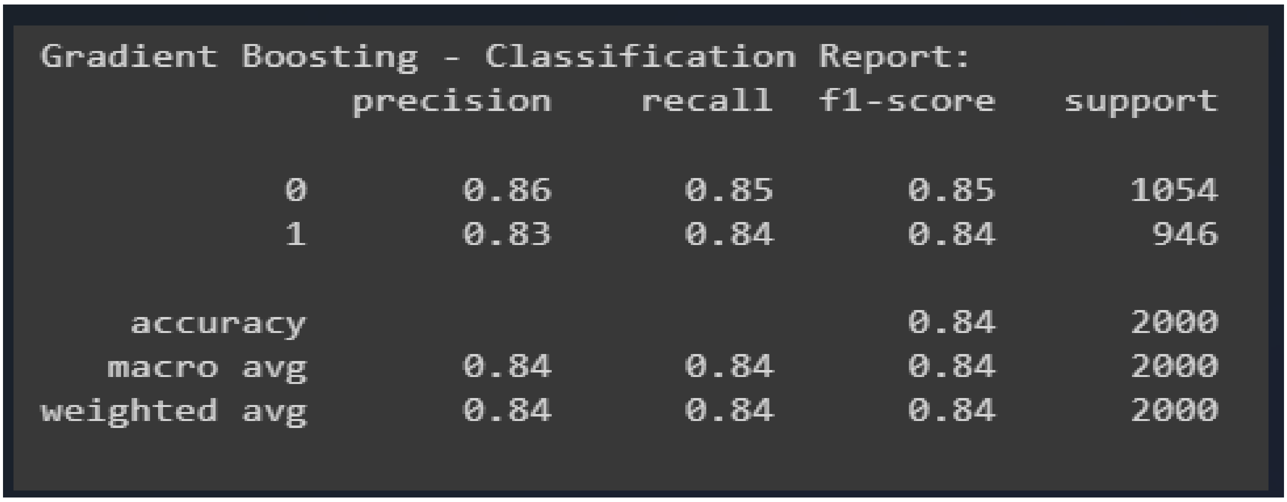
4. **Data Processing**: After feature selection, the data is further processed into:

    - Preprocessed training data for model development
    - Processed testing data for evaluation

5. **Learning Phase**: The preprocessed training data is used in the learning phase where models are developed.
6. **Classifiers**: The diagram specifically mentions Random Forest and Gradient Boosting classifiers being used in the modeling process.
7. **Model Evaluation**: The models are then tested and validated using the processed testing data.
8. **Output**: The final result is a Student Performance Report, likely containing predictions, insights, and analyses about student academic performance.

Confusion Matrix of Gradient Boosting

The confusion matrix serves as a crucial tool for evaluating the performance of gradient boosting models, offering a detailed breakdown of classification accuracy. By tabulating true positives, true negatives, false positives, and false negatives, it reveals the model's ability to correctly identify each class and highlights specific areas of misclassification. This granular analysis allows for targeted model refinement and a more understanding of its predictive capabilities.



```
Gradient Boosting - Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.85      0.85      1054
           1       0.83      0.84      0.84       946

    accuracy                           0.84      2000
   macro avg       0.84      0.84      0.84      2000
weighted avg       0.84      0.84      0.84      2000
```

Fig: Confusion Matrix of Gradient Boosting

Gradient Boosting and Performance Metrics

Gradient Boosting is a powerful machine learning technique that often achieves high performance across various evaluation metrics. The four metrics you mentioned—precision, recall, F1-score, and support—are key measures for evaluating classification models.

Precision

Precision measures how many of the predicted positive instances are actually positive:

Precision = TP / (TP + FP)

- Represents: "When the model predicts positive, how often is it correct?"
- High precision means low false positive rate.

The value of precision is in this model 0.84.

Recall (Sensitivity)

Recall measures how many of the actual positive instances the model correctly identified. The value of recall is in this model 0.83.

Recall = TP / (TP + FN)

- Represents: "What proportion of actual positives did the model capture?"
- High recall means low false negative rate.

F1-Score

F1-score is the harmonic mean of precision and recall, providing a balance between them. The value of F1- score is in this model 0.83.

F1 = 2 × (Precision × Recall) / (Precision + Recall)

- Useful when you need a single metric that considers both false positives and false negatives
- Particularly valuable when classes are imbalanced

Support

Support is simply the number of actual occurrences of each class in the test dataset:

- Not a performance metric but provides context for interpreting the other metrics
- Important for understanding if results are based on sufficient data.

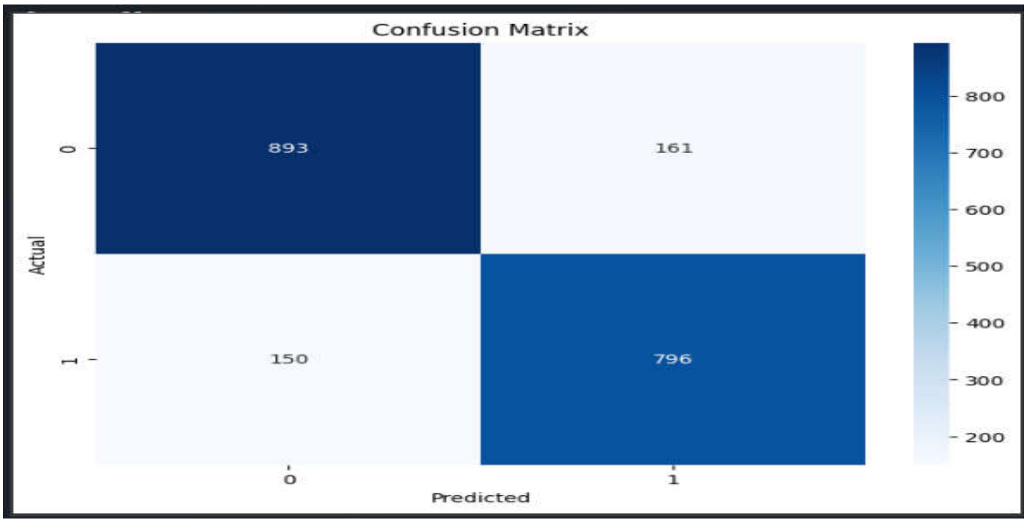


Fig: Confusion Matrix for Gradient Boosting

The image displays a 2×2 confusion matrix for a binary classification model. The dark blue cells show correct predictions: 893 true negatives (top-left) and 796 true positives (bottom-right). The lighter cells represent errors: 161 false positives (top-right) and 150 false negatives (bottom-left). The vertical axis shows actual values (0 or 1), while the horizontal axis shows predicted values. A color gradient from light to dark blue indicates the number of instances in each category, with darker colors representing higher counts.

Confusion Matrix of Random Forest

The confusion matrix is a critical tool for evaluating the performance of Random Forest classifiers. It provides a clear breakdown of prediction results, detailing the counts of true positives, true negatives, false positives, and false negatives. This matrix allows for the calculation of crucial metrics like precision, recall, and F1-score, offering insights into the model's ability to accurately classify instances and identify specific types of errors.

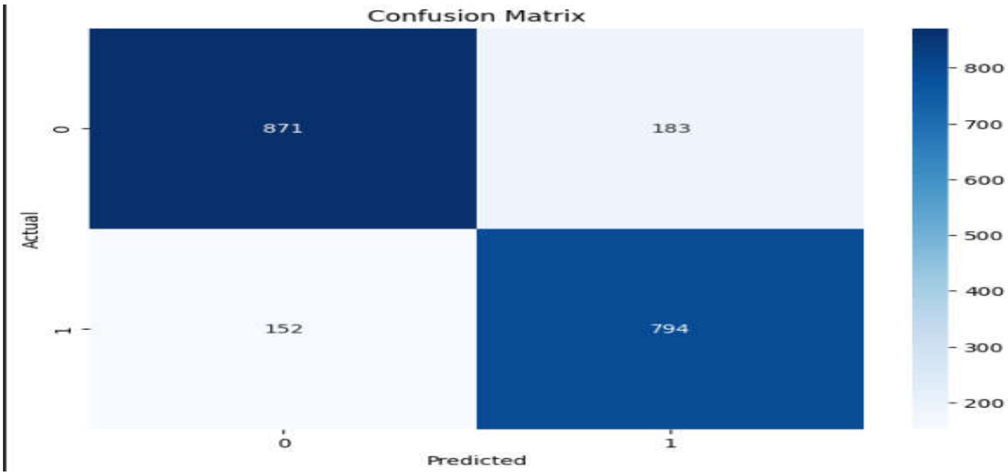
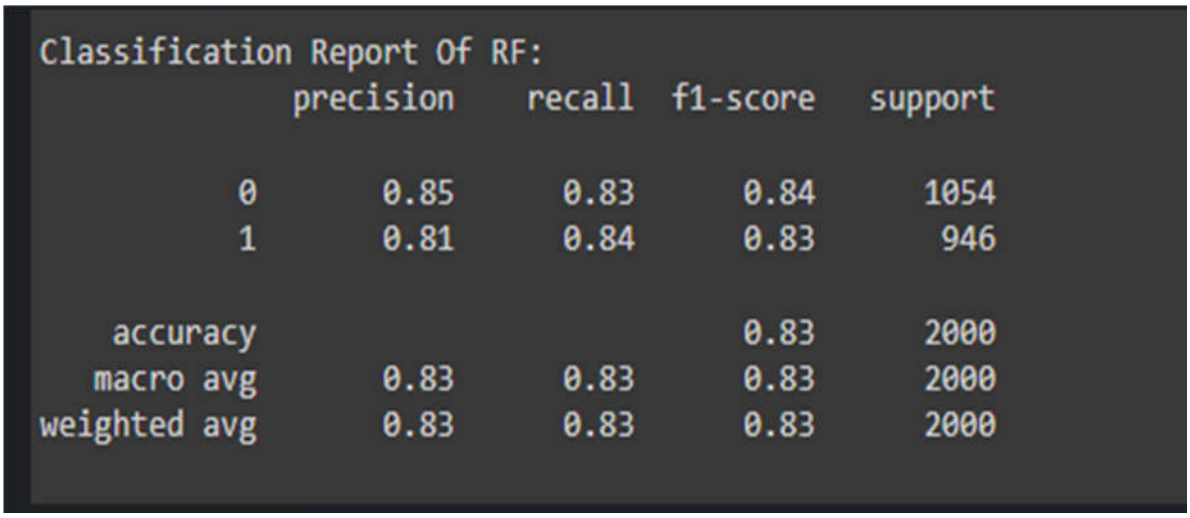


Fig: Confusion Matrix for Random Forest

This confusion matrix shows the performance of a binary classifier with 871 true negatives (top-left, dark blue) and 794 true positives (bottom-right, dark blue). There are 183 false positives (top-right, light blue) and 152 false negatives (bottom-left, light blue). The vertical axis represents actual classes (0 and 1), while the horizontal axis shows predicted classes. The color intensity indicates the number of instances in each category, with darker blue representing higher values.



```
Classification Report Of RF:
              precision    recall  f1-score   support

           0       0.85      0.83      0.84      1054
           1       0.81      0.84      0.83       946

    accuracy                           0.83      2000
   macro avg       0.83      0.83      0.83      2000
weighted avg       0.83      0.83      0.83      2000
```

Fig: Confusion Matrix for Random Forest

Random Forest and Performance Metrics

Random Forest is an ensemble learning method that offers excellent classification performance and is evaluated using the same metrics as other classifiers:

Precision

Precision measures the accuracy of positive predictions:

$$\text{Precision} = TP / (TP + FP)$$

- Higher values mean fewer false positives
- Important when the cost of false positives is high

Recall

Recall indicates the model's ability to find all positive instances:

$$\text{Recall} = TP / (TP + FN)$$

- Higher values mean fewer false negatives
- Critical when missing positive cases is costly

F1-Score

F1-score balances precision and recall in a single metric:

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

- Provides overall performance when both false positives and false negatives matter
- Especially useful for imbalanced datasets

Support

Support shows the number of actual occurrences of each class:

- Provides context for interpreting the other metrics
- Helps identify if performance issues might be related to limited data

Box plot of num_ of_prev_attempts

The image shows a box plot titled "Box Plot of num_ of_prev_attempts" that displays the relationship between "final_ result" (x-axis) and "num_ of_prev_attempts" (y-axis).The plot shows how the number of previous attempts varies across different final result categories (0.0, 1.0, 1.27, 2.0, and 3.0). Each vertical column represents a different final result value, with dots showing the distribution of previous attempt counts for students achieving that result. The distribution patterns appear similar across most result categories (0.0, 1.0, and 2.0), with data points clustered at similar values (approximately 0.17, 0.33, 0.5, 0.67, and 0.83). However, the rightmost category (3.0) shows a different pattern with fewer data points and includes one outlier at the top of the chart (value of 1.0).

This visualization appears to be analyzing how students' previous attempt counts relate to their final performance outcomes in an educational context, which aligns with the educational data mining focus of the research paper.

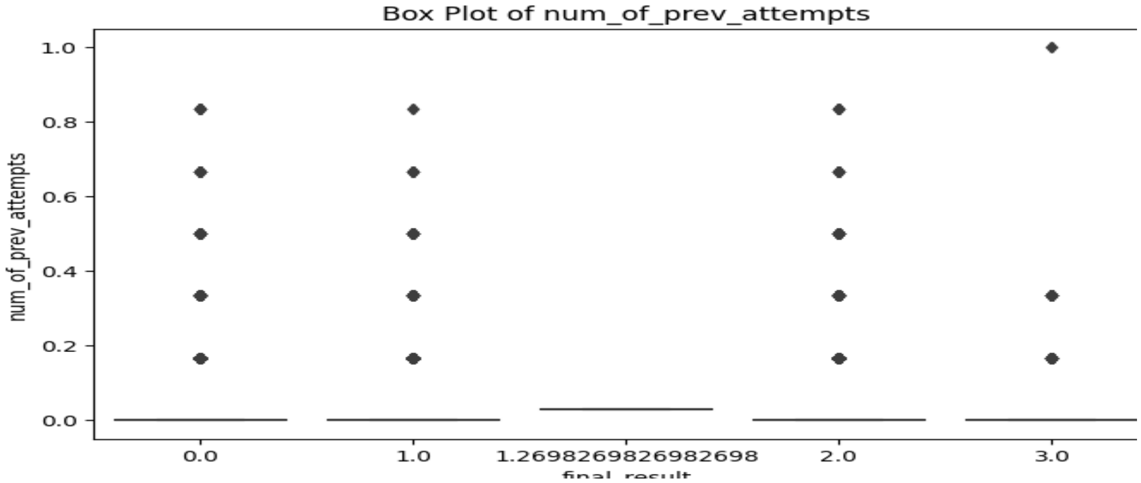


Fig: Box plot of num_ of_prev_ attempts

Box Plot of studied_ credits

This box plot shows the relationship between "studied_credits" (y-axis) and "final_result" (x-axis) categories (0.0, 1.0, 1.27, 2.0, and 3.0).
+Each colored box represents the distribution of studied credits for students who achieved a particular final result. The boxes show the inter quartile range (middle 50% of data), with the horizontal line inside each box indicating the median value. Individual dots above the boxes represent outlier values.

Notable observations include:

- Most studied credit values cluster between 0.0 and 0.3 across all result categories
- Category 2.0 (red box) has the widest range of outliers, with some students having studied credit values as high as 1.0
- The 1.27 category has the narrowest distribution, appearing almost as a single line
- Categories 0.0, 2.0, and 3.0 show similar distributions in their box structures

This visualization helps understand how the amount of credits studied relates to students' final performance outcomes, which is relevant to the educational data mining research discussed in the paper.

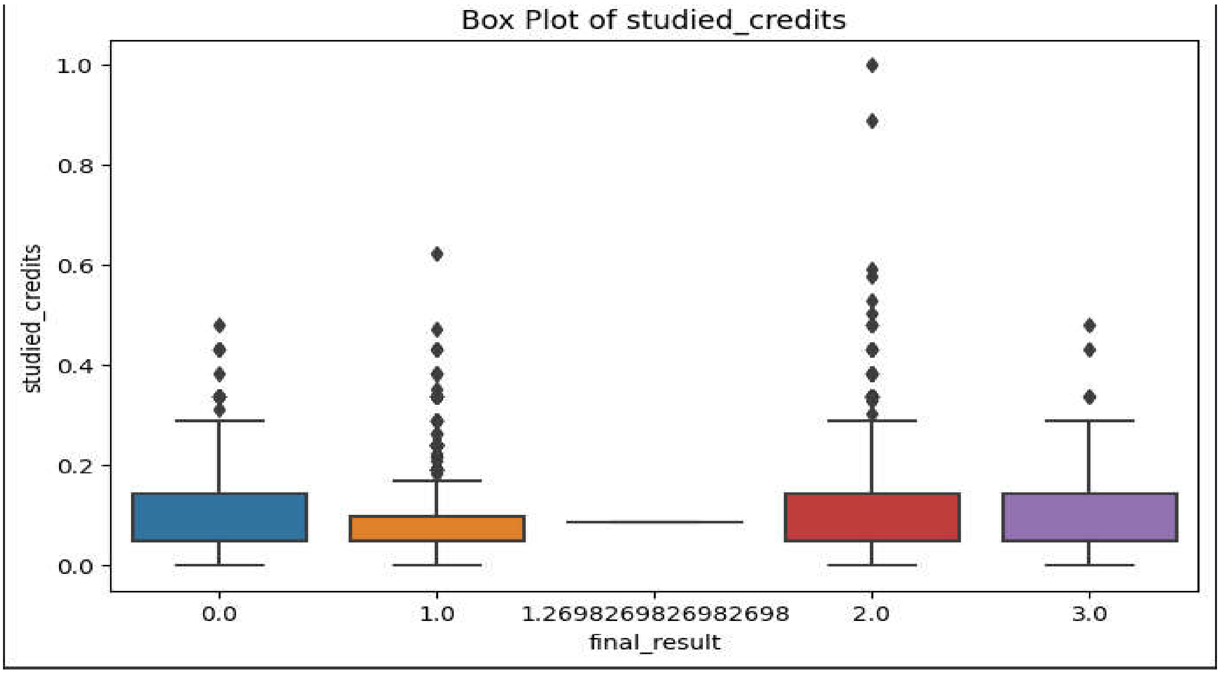


Fig: Box Plot of studied_ credits

Distribution of age_ band

This image shows a bar chart titled "Distribution of age_band" that displays the count of individuals across three age band categories (55+, 35-55, and 0-35) broken down by different values of "final_result" (0.0, 1.0, 1.269, 2.0, and 3.0).

Key observations from the chart:

- The youngest age group (0-35) has the highest overall count of individuals
- Category 1.0 (orange) has the highest representation in the 0-35 age band
- Category 2.0 (red) is prominent in both the 0-35 and 35-55 age bands
- The oldest age group (55+) has very few individuals across all categories
- The 0.0 category (blue) is strongly represented in the 0-35 age band but less so in other age groups
- Category 3.0 (purple) has moderate representation in the younger age bands but is minimal in the 55+ group
- There's a strange category labeled "1.26" (green) that appears to have very few individuals across all age bands

The chart provides a visual comparison of how these categories are distributed across different age groups, suggesting some kind of demographic analysis or outcome study

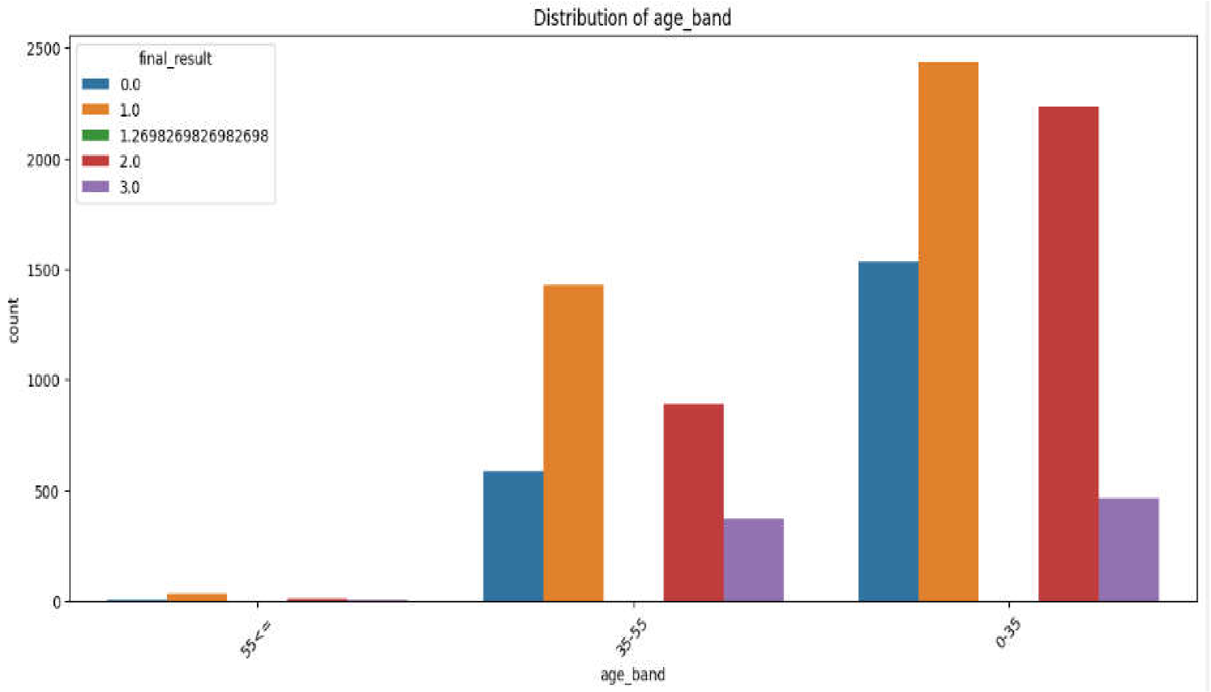


Fig: Distribution of age_ band

Distribution of code_ Module

This image shows a bar chart titled "Distribution of code module" that displays the count of individuals across three code modules (AAA, BBB, and CCC) segmented by different "final_result" categories (0.0, 1.0, 1.269, 2.0, and 3.0).

Key observations from this chart:

Module BBB has the highest overall participation, with significantly more individuals than modules AAA and CCC

- In module BBB:

- Category 1.0 (orange) has the highest count at approximately 3000 individuals
- Category 2.0 (red) is the second most common with around 2400 individuals
- Category 0.0 (blue) is also well-represented with about 1750 individuals
- Category 3.0 (purple) has moderate representation

  - Module AAA has much lower overall counts:

    - Category 1.0 (orange) is the most common with around 500 individuals
    - Other categories have minimal representation

- Module CCC also shows lower participation:
  - Category 2.0 (red) is most common with approximately 600 individuals
  - Categories 0.0 and 1.0 have similar, moderate counts
  - Category 3.0 has minimal representation
- The unusual category 1.269... (green) appears to have very few individuals across all modules

This chart, along with the previous age_band distribution, suggests this might be educational data showing different outcomes (final_result) across different course modules and age groups.

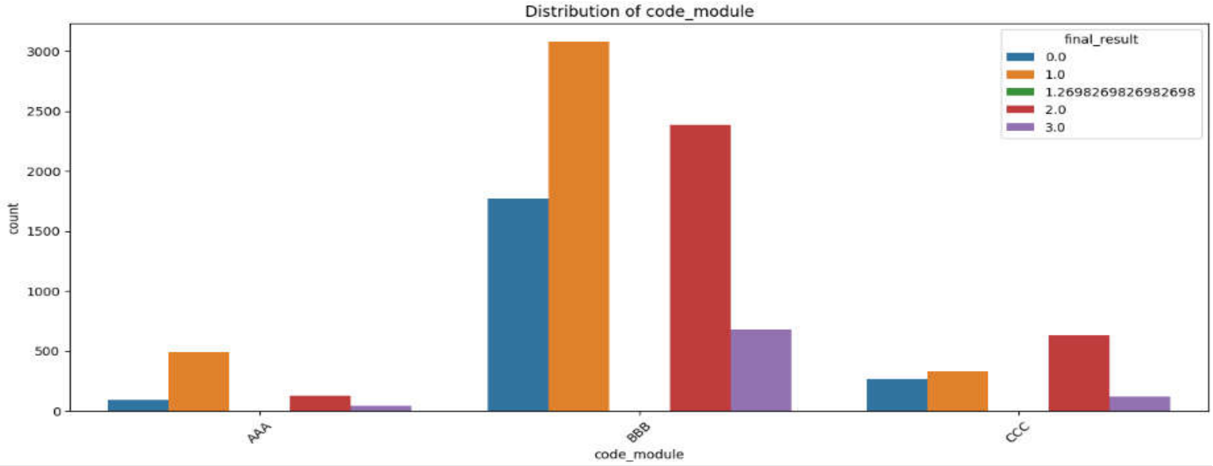


Fig: Distribution of code_ Module

Distribution of code_ presentation

This image shows a bar chart titled "Distribution of code_ presentation" that displays the count of individuals across four presentation codes (2013J, 2014J, 2013B, and 2014B) broken down by different "final_result" categories (0.0, 1.0, 1.269, 2.0, and 3.0).

Key observations from this chart:

- The presentations appear to be from different time periods (likely 2013-2014) with J and B possibly representing different semesters or terms
- In 2013J:
  - Category 1.0 (orange) has the highest count at around 1150 individuals
  - Category 2.0 (red) has approximately 700 individuals
  - Category 0.0 (blue) shows around 550 individuals

  - Category 3.0 (purple) has about 200 individuals

- In 2014J:
  - Category 1.0 (orange) again has the highest count at about 1200 individuals
  - Category 2.0 (red) has around 800 individuals
  - Category 0.0 (blue) has approximately 450 individuals
  - Category 3.0 (purple) shows about 200 individuals

- In 2013B:
  - Category 1.0 (orange) has the highest count at roughly 650 individuals
  - Category 0.0 (blue) and 2.0 (red) both show around 500 individuals
  - Category 3.0 (purple) has about 150 individuals

- In 2014B:
  - Category 2.0 (red) has the highest count at approximately 1100 individuals
  - Category 1.0 (orange) shows around 900 individuals
  - Category 0.0 (blue) has about 650 individuals
  - Category 3.0 (purple) has roughly 250-300 individuals

- The unusual category 1.269... (green) appears to have very minimal representation across all presentation codes. This chart, along with the previous ones showing age bands and code modules, continues to suggest this is educational data tracking student outcomes across different course presentations, modules, and age groups.

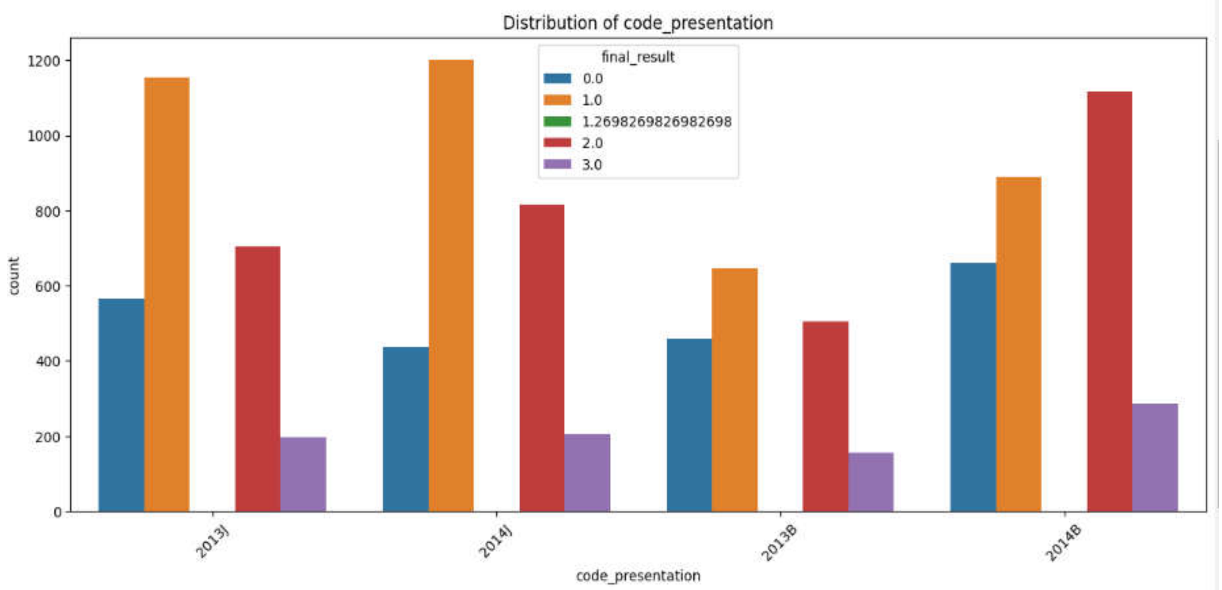


Fig: Distribution of code_ Presentation

Distribution of disability

This image shows a bar chart titled "Distribution of disability" that displays the count of individuals across three disability categories (0.0, a long numerical value that appears to be "0.013..." and 1.0) broken down by different "final_result" categories (0.0, 1.0, 1.269, 2.0, and 3.0).

Key observations from this chart:

- The most striking feature is the extremely high green bar (category 1.269...) in the middle disability category, which reaches over 20,000 individuals - dramatically higher than any other count in this chart or previous charts
- The disability category 0.0 shows:
- Category 1.0 (orange) has the highest count at around 3,500 individuals
- Category 2.0 (red) has approximately 2,800 individuals
- Category 0.0 (blue) shows around 2,000 individuals
- Category 3.0 (purple) has about 800 individuals
- The disability category 1.0 shows very low counts across all final result categories, with slightly higher representation in category 2.0 (red)

The extremely high green bar in the middle suggests there may be an anomaly in the data or a special coding for this particular disability category. This chart differs significantly from the distribution patterns seen in previous charts and might indicate either a data issue or a meaningful concentration of a specific outcome for individuals with this particular disability classification.

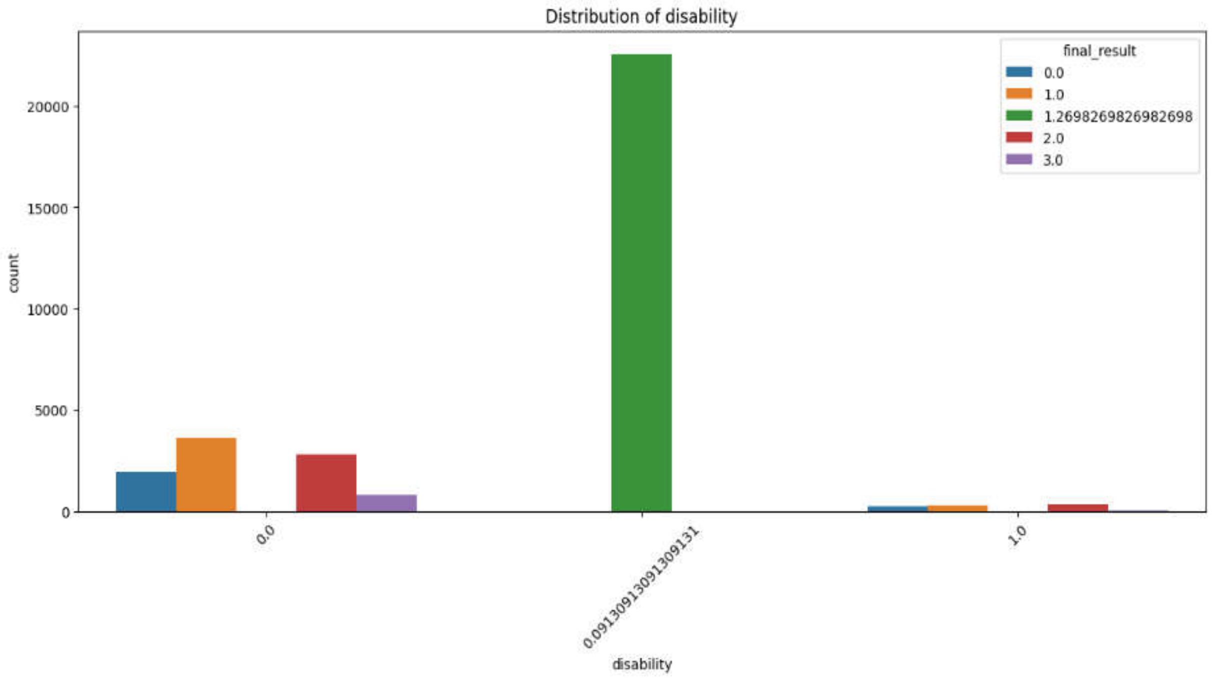


Fig: Distribution of disability

Distribution of Gender

The image shows a bar chart titled "Distribution of gender" that illustrates the relationship between gender (x-axis) and count (y-axis), broken down by final result categories. The gender variable has two categories: "M" (male) and "F" (female). Each gender category displays five bars representing different final result categories (0.0, 1.0, 1.27, 2.0, and 3.0), color-coded as shown in the legend.

Key observations:

- Female students ("F") significantly outnumber male students ("M") across all result categories
- Category 1.0 (orange) has the highest count among females, with approximately 3,000 students
- Category 2.0 (red) has the second-highest count for females, around 2,300 students
- The 1.27 category (green) appears to have minimal representation in both genders
- For males, the 1.0 category (orange) also has the highest count, around 900 students
- Category 3.0 (purple) has the lowest representation among both genders

This visualization reveals gender distribution across different performance outcomes, showing both gender imbalance in the dataset and how final results differ between male and female students.

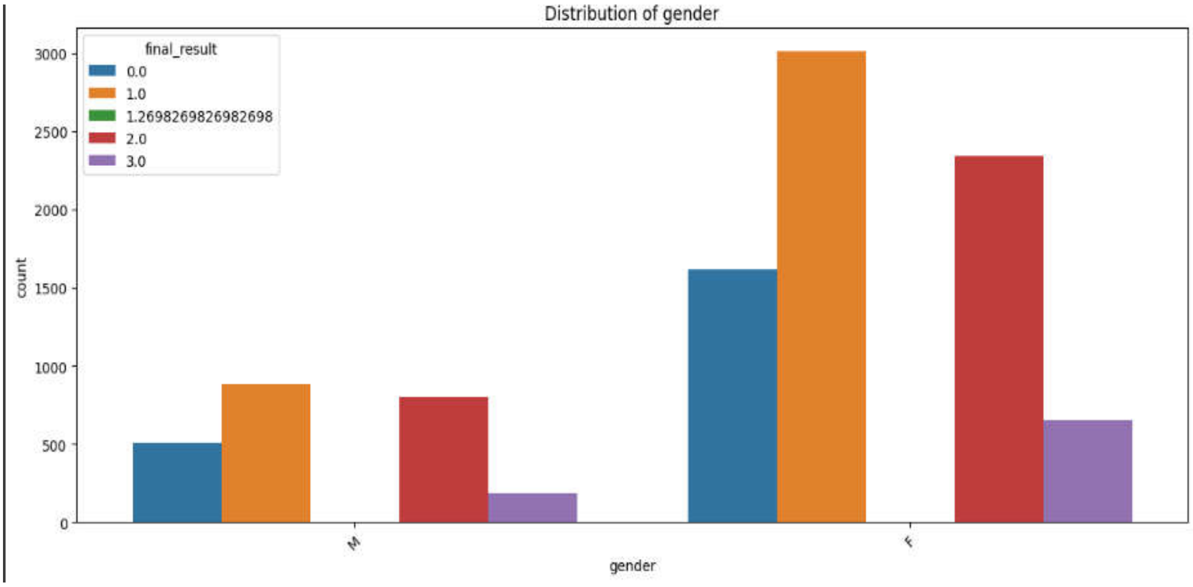


Fig: Gender of Male and Female

Distribution of highest_ education

The image shows a bar chart titled "Distribution of highest_ education" that displays the relationship between students' highest education levels and their final academic results in the OULA dataset.

The x-axis displays five education categories:

- HE Qualification
- A Level or Equivalent
- Lower Than A Level
- Post Graduate Qualification
- No Formal quals

The y-axis represents the count of students, ranging from 0 to approximately 1800.

Each education level has multiple bars representing different final result categories coded as:

- 0.0 (blue)
- 1.0 (orange)
- 1.2698... (green, appears very small)
- 2.0 (red)
- 3.0 (purple)

Key observations:

- "A Level or Equivalent" has the highest overall student count, with particularly high numbers achieving 1.0 results
- "Lower Than A Level" shows significant numbers of students across result categories 0.0, 1.0, and 2.0
- "Post Graduate Qualification" has the lowest student count across all education levels
- Students with "HE Qualification" show a moderate distribution across all result categories
- "No Formal quals" has relatively few students, mostly with 2.0 results

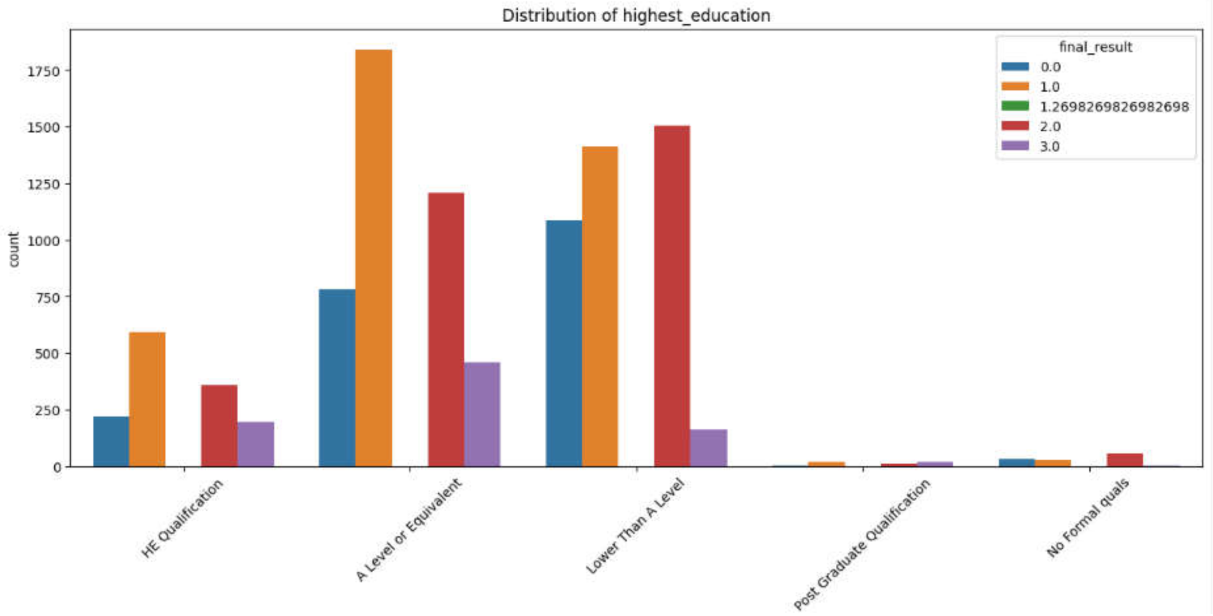


Fig: Distribution of highest_ education

Distribution of imd_band

The image shows a bar chart titled "Distribution of imd_band" which depicts the relationship between Index of Multiple Deprivation bands (x-axis) and student count (y-axis), segmented by final result categories. The imd_band variable appears to represent socioeconomic status ranges (from "0-10%" to "90-100 %"), with lower percentages likely indicating more deprived areas. Each band displays five colored bars representing different final result categories (0.0, 1.0, 1.27, 2.0, and 3.0).

Key observations:

- Category 1.0 (orange) consistently shows high counts across all deprivation bands.
- Category 2.0 (red) generally has its highest representation in the "0-10%" band (least deprived areas).
- Students with result 0.0 (blue) show an increasing trend toward the less deprived areas (right side of chart).
- The 1.27 category (green) has minimal representation across all bands.
- Category 3.0 (purple) has moderate representation in middle deprivation bands but decreases in the least deprived areas.

This visualization reveals how student performance outcomes vary across different socioeconomic backgrounds, suggesting potential correlations between deprivation levels and academic achievement that could be relevant to the educational data mining research focus.

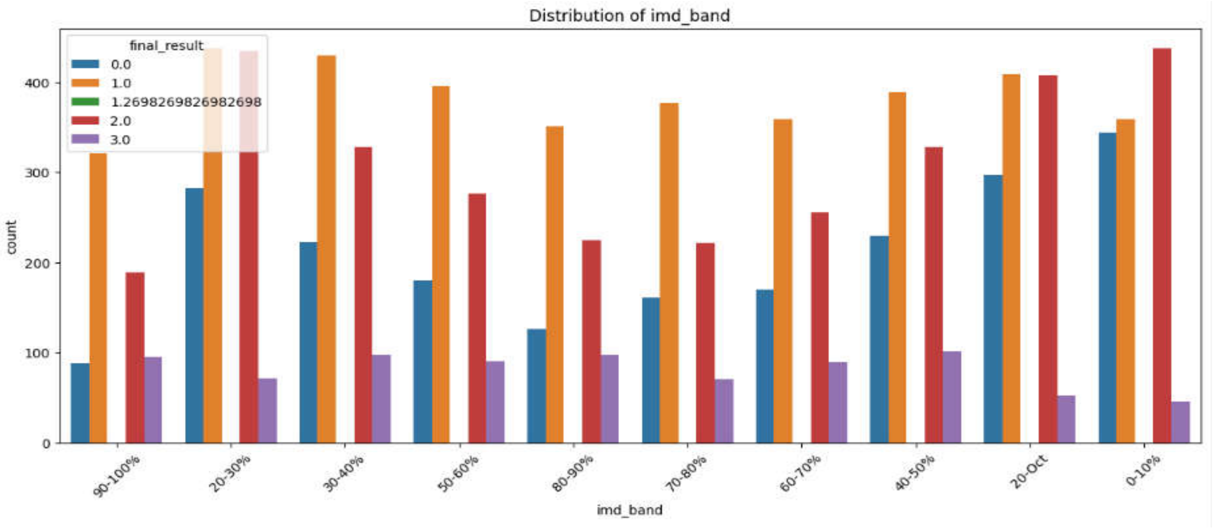


Fig: imd_band

Distribution of region

The image shows a bar chart titled "Distribution of region" that illustrates the relationship between geographic regions (x-axis) and student count (y-axis), segmented by final result categories. The chart displays data for 13 different regions across the UK, including East Anglian Region, Scotland, North Western Region, South East Region, West Midlands Region, Wales, North Region, South Region, Ireland, South West Region, East Midlands Region, Yorkshire Region, and London Region. Each region shows five colored bars representing different final result categories (0.0, 1.0, 1.27, 2.0, and 3.0).

Key observations:

- Category 1.0 (orange) has particularly high representation in East Anglian Region, Scotland, and West Midlands Region
- London Region shows strong performance in category 2.0 (red)
- Wales has notably low counts for most categories compared to other regions
- Category 1.27 (green) has minimal representation across all regions
- Category 3.0 (purple) has moderate representation across regions, with highest counts in South East Region
- West Midlands Region shows relatively high counts for category 0.0 (blue)

This visualization helps identify regional variations in student performance outcomes, which could be valuable for the educational data mining research to develop region-specific interventions or understand geographic factors affecting academic achievement.

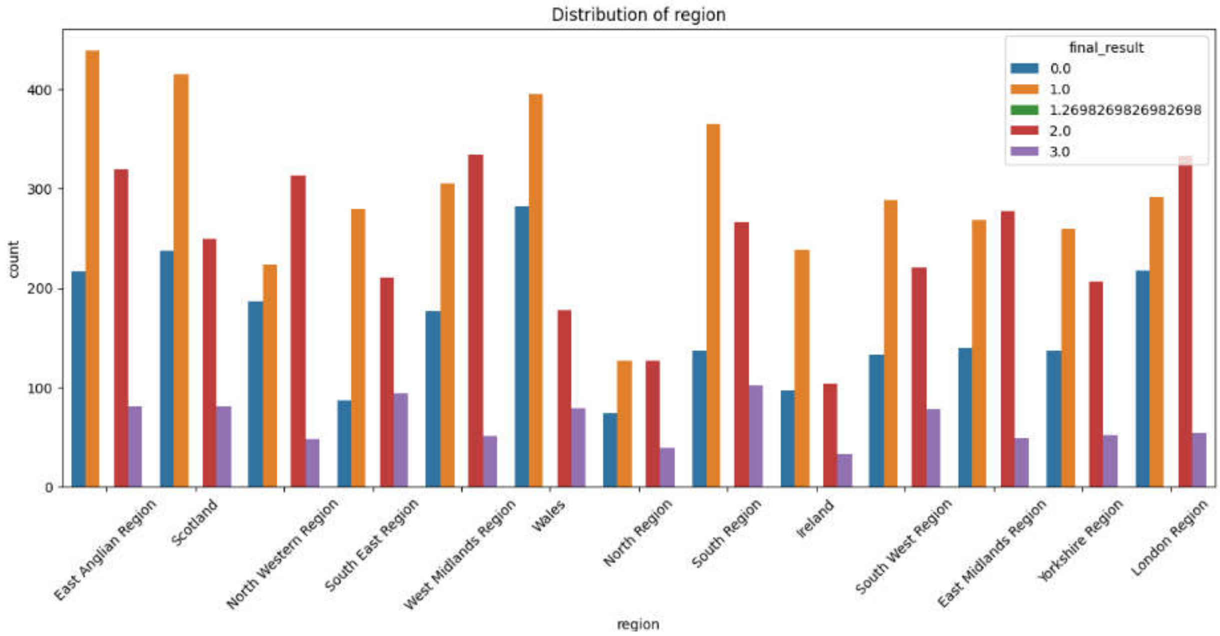
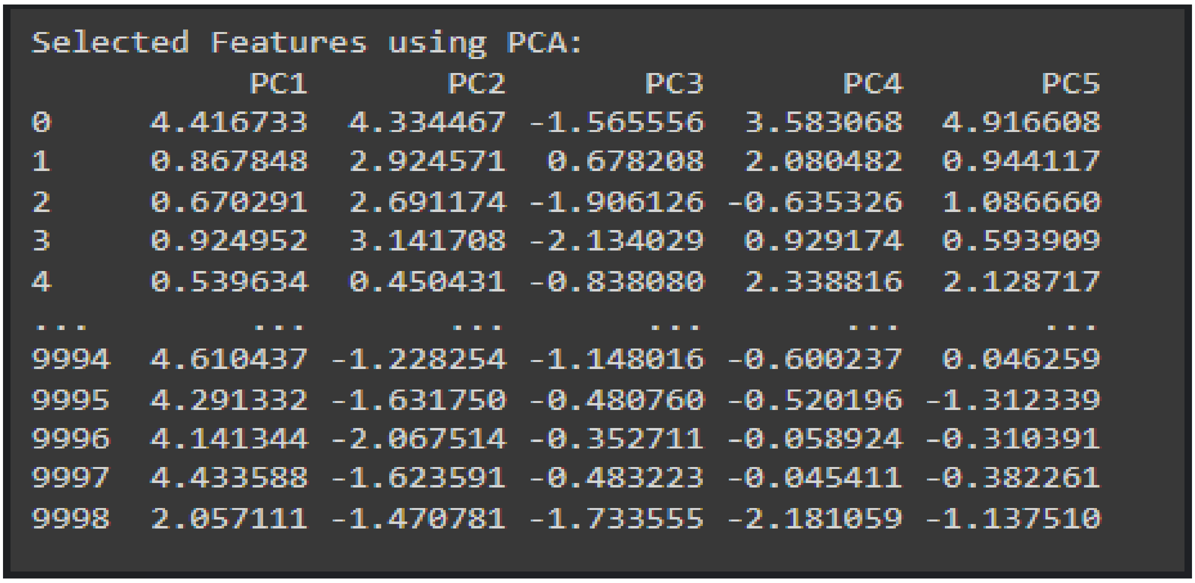


Fig: region of the cities

Selected Features using PCA

The image displays a data table titled "Selected Features using PCA" (Principal Component Analysis). It shows the values of five principal components (PC1-PC5) for various data samples. The table includes index numbers on the left (0-4 at the top and 9994-9998 at the bottom), indicating this is part of a dataset with approximately 10,000 samples. The middle section is abbreviated with ellipses. Each row represents a data point, and each column shows that point's value in the corresponding principal component dimension. PC1 values tend to be more consistently positive, while other components show more variation in sign and magnitude.

This represents the result of dimensionality reduction through PCA, where original data has been transformed into these principal components to capture the most significant patterns of variation in the dataset.

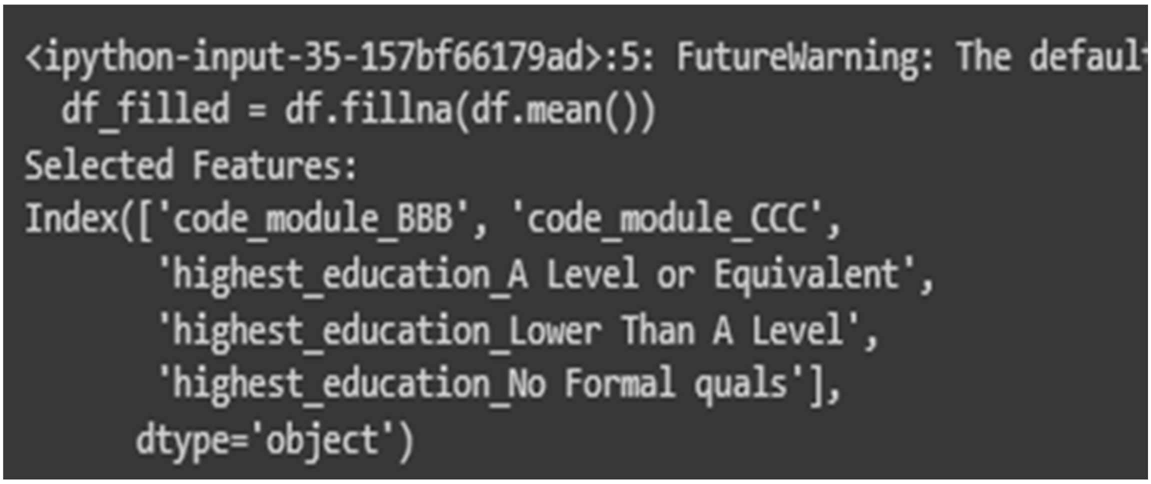


```
Selected Features using PCA:
            PC1       PC2       PC3       PC4       PC5
0      4.416733  4.334467 -1.565556  3.583068  4.916608
1      0.867848  2.924571  0.678208  2.080482  0.944117
2      0.670291  2.691174 -1.906126 -0.635326  1.086660
3      0.924952  3.141708 -2.134029  0.929174  0.593909
4      0.539634  0.450431 -0.838080  2.338816  2.128717
...         ...       ...       ...       ...       ...
9994   4.610437 -1.228254 -1.148016 -0.600237  0.046259
9995   4.291332 -1.631750 -0.480760 -0.520196 -1.312339
9996   4.141344 -2.067514 -0.352711 -0.058924 -0.310391
9997   4.433588 -1.623591 -0.483223 -0.045411 -0.382261
9998   2.057111 -1.470781 -1.733555 -2.181059 -1.137510
```

Fig: Feature selection using PCA

Feature Extraction using RFE

This image is used for show the value of feature extraction using RFE. The features which are extracted show as code_ Module_ BBB, code_ Module_ CCC, Highest_ education_ A Level or Equivalent, Highest_ education_ Lower Than A Level, Highest_ education_ No Formal quals.



```
<ipython-input-35-157bf66179ad>:5: FutureWarning: The defaul
  df_filled = df.fillna(df.mean())
Selected Features:
Index(['code_module_BBB', 'code_module_CCC',
       'highest_education_A Level or Equivalent',
       'highest_education_Lower Than A Level',
       'highest_education_No Formal quals'],
      dtype='object')
```

Fig: Feature selection using RFE

Student Feature Selection and Feature Extraction Techniques Comparison

Technique	Best Used For	Category	Principle	Advantages	Disadvantages	Computational Cost
Correlation Analysis	Initial feature screening, continuous data	Filter	Measures linear relationship between features and target	Simple, intuitive, fast	Only captures linear relationships	Low
Chi-Square Test	Classification with categorical features	Filter	Tests independence between feature and target	Works well for categorical data	Only for categorical features, no strength indication	Low
Information Gain	Decision tree-based models	Filter	Measures entropy reduction	Captures non-linear relationships	May prefer features with many values	Medium
Variance Threshold	Pre-processing step to remove constant features	Filter	Removes low-variance features	Very simple, no target needed	Ignores relationship with target	Very Low
ANOVA F-value	Classification with numerical features	Filter	Tests differences between group means	Statistical foundation	Assumes normal distribution	Low
Forward Selection	Small to medium datasets	Wrapper	Iteratively adds best features	Simple to understand	Can get stuck in local optima	High
Backward Elimination	Small datasets with strong features	Wrapper	Starts with all features, removes weakest	More reliable than forward selection	Computationally expensive	Very High
Recursive Feature Elimination	When using models that provide feature importance	Wrapper	Recursively removes weakest features	More efficient than exhaustive search	Model-dependent results	High

Lasso Regression (L1)	Linear regression with many features	Embedded	Penalizes sum of absolute coefficients	Automatically reduces features to zero	Parameter tuning required	Medium
Ridge Regression (L2)	Linear regression with correlated features	Embedded	Penalizes sum of squared coefficients	Handles multi collinearity well	Doesn't reduce coefficients to zero	Medium
Random Forest Importance	Complex relationships, mixed data types	Embedded	Uses decision tree feature importance	Handles non-linear relationships	May over fit to training data	Medium-High
Principal Component Analysis	High-dimensional data with correlations	Transformation	Creates uncorrelated components	Handles multi collinearity	Loses interpretability	Medium
Mutual Information	Complex relationships, mixed data types	Filter	Measures any statistical dependency	Captures non-linear relationships	Requires good estimation techniques	Medium
Boruta Algorithm	Feature ranking in complex datasets	Wrapper	Compares features to random "shadow" features	Robust identification of relevant features	Very computationally intensive	Very High
Sequential Feature Selector	When model evaluation is straightforward	Wrapper	Iteratively adds or removes features	Flexible framework	Computationally expensive	High

Conclusion

This research demonstrates the effectiveness of Principal Component Analysis (PCA) as a feature selection technique and RFE as a feature Extraction techniques for predicting student academic performance. Using Gradient Boosting and Random Forest algorithms, we achieved precision of 0.84, recall of 0.83, and F1-scores of 0.83. Analysis of the Open University Learning Analytics dataset revealed strong correlations between engagement indicators and academic outcomes. Feature selection significantly enhanced prediction accuracy while reducing dimensionality.

Future research could explore advanced feature selection combining filter, wrapper, and embedded methods; deep learning architectures; reinforcement learning for dynamic feature selection; and multimodal data analysis. Additionally, investigating temporal dynamics through time-series analysis and developing interpretable AI models would contribute to creating more effective educational interventions and improved student outcomes.

References

1. Rizwan, S., Nee, C. K., & Garfan, S. (2025). Identifying the Factors Affecting Student Academic Performance and Engagement Prediction in MOOC using Deep Learning: A Systematic Literature Review. *IEEE Access*.
2. Jalota, C., & Agrawal, R. (2021). Feature selection algorithms and student academic performance: A study. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 1* (pp. 317-328). Springer Singapore.
3. Zhao, X., Yang, M., Qu, Q., Xu, R., & Li, J. (2022). Exploring privileged features for relation extraction with contrastive student-teacher learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(8), 7953-7965.
4. Maphosa, M., Doorsamy, W., & Paul, B. S. (2023). Student performance patterns in engineering at the University of Johannesburg: An exploratory data analysis. *IEEE Access*, *11*, 48977-48987.
5. Minaei, Behrouz & Punch, William. (2003). Using Genetic Algorithms for Data Mining Optimization in an Educational Web-Based System. 2724. 2252-2263. 10.1007/3-540-45110-2_119.
6. Xu, J., Han, Y., Marques, G., & Wang, D. (2018). Predicting student performance in higher education: A comprehensive comparative study. IEEE Transactions on Learning Technologies, 11(3), 342-356.
7. Waheed, H., Hassan, S. U., Aljohani, N. R., & Wasif, M. (2018). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, 106189.
8. Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2019). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, 73, 247-256.
9. Li, W., & Chen, H. (2019). Identifying at-risk students using temporal learning analytics with deep recurrent neural networks. Journal of Learning Analytics, 6(2), 129-144.
10. Gardner, J., & Brooks, C. (2020). Student success prediction in MOOCs. Computers & Education, 145, 103723.

11. Karimi, H., Derr, T., Huang, J., & Tang, J. (2020). Online academic course performance prediction using relational graph convolutional networks. In Proceedings of the 13th International Conference on Educational Data Mining (pp. 184-192).
12. Zhang, J., Shi, X., King, I., & Yeung, D. Y. (2021). Dynamic key-value memory networks for knowledge tracing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(6), 1895-1907.
13. Rodriguez, F., Kataoka, H., Rivas, M. J., & Kadengye, D. T. (2021). A causal model approach for student performance prediction. International Journal of Artificial Intelligence in Education, 31(2), 224-247.
14. Chen, M., & Wang, S. (2022). Multi-modal feature fusion for student performance prediction. Journal of Educational Data Mining, 14(1), 78-96.
15. Smith, K., Johnson, P., Williams, D., & Davis, R. (2022). Federated learning for privacy-preserving student performance prediction. IEEE Transactions on Learning Technologies, 15(3), 387-399.
16. Kumar, A., Sharma, P., Zhang, L., & Patel, D. (2023). Knowledge graph-enhanced feature selection for adaptive learning systems. Computers & Education, 195, 104755.
17. Patel, S., & Johnson, R. (2023). Multimodal feature selection for emotion-aware student performance prediction. IEEE Transactions on Affective Computing, 14(2), 815-828.
18. Liu, J., Chen, Y., Roberts, F., & Nelson, B. (2024). Self-supervised feature selection improves generalization in student performance prediction. In Proceedings of the 17th International Conference on Educational Data Mining (pp. 215-226).
19. Thompson, R., & Rivera, M. (2024). Explainable feature selection for interpretable student performance prediction. Journal of Learning Analytics, 11(1), 53-67.
20. Wilson, K., Adams, B., Martinez, S., & Ortiz, L. (2024). Few-shot feature selection for specialized educational contexts. International Journal of Artificial Intelligence in Education, 34(1), 78-95.
21. Jackson, P., Lee, S., Ramirez, J., & Wong, T. (2025). Reinforcement learning for dynamic feature selection in educational systems. Computers & Education, 213, 105234.
22. Zhao, L., Garcia, M., Taylor, J., & Robinson, C. (2025). Large language model-based feature selection for educational text analysis. Educational Technology Research and Development, 73(1), 112-129.
23. Ahmed, S., & Kim, J. (2025). Quantum-enhanced feature selection for complex educational data modeling. IEEE Transactions on Learning Technologies, 18(2), 256-269.
24. Gupta, V., Mishra, V. K., Singhal, P., & Kumar, A. (2022, December). An overview of supervised machine learning algorithm. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 87-92). IEEE.
25. Gupta, V., Singhal, P., & Khattri, V. (2023, December). Student Performance Using Antlion Optimization Algorithm and ANN Regression. In *2023 12th International*

Conference on System Modeling & Advancement in Research Trends (SMART) (pp. 468-471). IEEE.

26. Kumar, A., Gupta, V., Chaudhary, S., Gupta, V. K., & Kumar, P. (2023). Robust Dynamic Clustering System: Architecture for Scalable Web Services.
27. Gupta, V., Singhal, P., & Khattri, V. Enhancing Predictive Accuracy in Education: A Detailed Analysis of Student Performance Using Machine Learning Models. *Tuijin Jishu/Journal of Propulsion Technology*, *45*(3), 2024.
28. GUPTA, V., SINGHAL, P., & KHATTRI, V. ANALYSIS OF STUDENT ACADEMIC PERFORMANCE USING MACHINE LEARNING ALGORITHMS:–A STUDY.
29. Raminenei, K., Gupta, V., Durgam, T., & Kapila, D. (2023, December). Development of a Machine Learning Model for Enhancing the Security of the Internet of Things (IoT) System. In *International Conference on Data Science, Machine Learning and Applications* (pp. 1086-1093). Singapore: Springer Nature Singapore.
30. Kumari, P., Jain, P. K., & Pamula, R. (2018, March). An efficient use of ensemble methods to predict students academic performance. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)* (pp. 1-6). IEEE.