Enhancing Dropout Prediction in Higher Education using a Hybrid Machine Learning Approach

Dr S Selvakani[1], K.Vasumathi[2]

[1]Assistant Professor and Head, PG Department of Computer Science, Government Arts and Science College, Arakkonam

[2]Assistant Professor, PG Department of Computer Science, Government Arts and Science College, Arakkonam

ABSTRACT:

Global student dropout is a significant problem that impacts not only the individual who leaves but also their college, family, and society as a whole. With the advancement of science and technology, big data has become a crucial tool for data analysis. Efficient prediction of student dropout from recorded educational data is currently a hot topic of research. The aim of this study was to investigate the factors that influence dropout rates among undergraduate students at Government Arts and Science College in Arakkonam.

To achieve this goal, a novel stacking ensemble combining Random Forest (RF) and Feed-forward Neural Networks (FNN) was proposed to predict dropout rates at the college. The proposed method was tested on a dataset collected from Government Arts and Science College in Arakkonam from 2016 to 2022. Results demonstrated that the proposed method outperformed base models in terms of testing accuracy and area under the curve (AUC) evaluation metrics.

To gather data, a combination of quantitative and qualitative methods was used. A survey was developed to determine the factors that influence student attrition from university programs. 100 students who had dropped out were contacted either over the phone or in person to explain the reasons behind their decision. The findings of the study indicate that the most significant factors contributing to student dropout were the pressure of studying, financial problems, and the desire to study abroad, which ranked second and third, respectively.

KEYWORDS: Student dropout, Machine learning, Random Forest, Neural Networks, Educational data

1. INTRODUCTION:

The issue of student dropout is widely recognized as a highly complex and significant problem within the education system. It not only causes economic, social, academic, political, and financial damage to all the key stakeholders involved in education, but also highlights the need for effective and efficient strategies to minimize the rate of undergraduate dropout. Despite measures taken in the past, the positive effects have not been evident. The most commonly used

definition of dropout focuses on whether a student continues to be active until the end of the course or the current semester.

Identifying students who are at risk of dropping out early on is crucial to reducing the problem and targeting the necessary conditions. Therefore, timing plays an important role in tackling the issue of dropout. Research has found that 75% of dropouts occur in the first few weeks, and predicting dropout is often considered a time series prediction or a sequence labeling problem. Alternatively, the time dimension can be indirectly incorporated into the prediction of dropout by using input features that are available within a specific time window. This allows for the selection of a suitable form of intervention. It is important to note that student dropout causes educational deficiencies, which can significantly affect the social and economic well-being of both current and future generations.

The issue of student dropout is significant as it can lead to economic, social, academic, political, and financial damage for everyone involved in the education system. The consequences of dropout can negatively impact the standards of living, employment opportunities, and cause disruptive behaviors in society. As a result, researchers, policymakers, and educators view dropout as a hindrance to educational development. To address this issue, early identification of at-risk students is critical, and a warning system can help colleges identify behaviors that may accelerate the risk of dropout and take proactive measures to prevent it from occurring.

Machine learning has shown promise in building predictive models for student dropout and providing early warning to authorities. Many studies have compared the performance of algorithms used in dropout prediction systems. However, this study aims to propose a novel approach based on a hybrid of Random Forest and Feed-forward Neural Networks to predict student dropout in university classes. By conducting a systematic review of the literature, the study seeks to integrate the reasons for dropout at different levels of education and create a conceptual framework to control dropout at the undergraduate and other levels of education in Tamilnadu.

The proposed study can help policymakers build an early warning system to signal which institutions need to control dropout rates. It can also benefit guardians of learners by providing insight into the internal and external reasons for dropout and their contribution to minimizing dropout at different levels of education.

The paper presents a novel contribution to educational data analysis by developing a stacking ensemble model combining Random Forest (RF) and Feed-forward Neural Networks (FNN) for predicting student dropout rates. The model, tested on data collected from 2016 to 2022, demonstrated superior accuracy and effectiveness compared to base models. Additionally,

the study identified key dropout factors, such as study pressure and financial problems, providing actionable insights for educational institutions. This research offers valuable policy recommendations for early intervention to reduce dropout rates, significantly impacting educational standards and economic progress in Tamilnadu.

The paper is well-structured, beginning with an abstract that summarizes the research problem, methodology, and key findings. The organization of the paper is structured as follows. **1. Abstract** provides a summary of the research problem, methodology, and key findings. **2. Introduction** covers the background of the study, the importance of addressing student dropout, and the objective of the study. In **3. Related Works**, previous research is reviewed, focusing on classification techniques in education and summarizing related works. **4. Research Methodology** details the data collection process, including surveys and interviews, describes the dataset, and outlines the machine learning models used. **5. Experimental Results and Discussion** explains the implementation of the proposed method, presents the performance metrics, discusses the findings, and acknowledges the study's limitations. **6. Contribution to the Work** highlights the novel stacking ensemble approach, the data collection and analysis, the identification of dropout factors, and provides policy recommendations. **7. Conclusion** summarizes the key findings, reiterates the policy recommendations, and suggests directions for future work. Finally, the paper includes **8. References**.

2. CONTRIBUTION TO THE WORK

Novel Stacking Ensemble Approach: The study introduces a unique stacking ensemble that combines Random Forest and Feed-forward Neural Networks to predict student dropout rates. This hybrid model shows improved accuracy and robustness compared to traditional models.

Data Collection and Analysis: The study uses a comprehensive dataset from Government Arts and Science College in Arakkonam, covering student data from 2016 to 2022. It employs both quantitative and qualitative methods to identify factors influencing dropout rates.

Identification of Dropout Factors: The research identifies critical factors contributing to student dropout, such as academic pressure, financial issues, and the desire to study abroad, providing a deeper understanding of the reasons behind student attrition.

Policy Recommendations: Based on the findings, the study proposes actionable recommendations for policymakers to develop strategies that address and mitigate dropout rates, thus enhancing the quality of education and economic progress.

3. RELATED WORKS:

To reduce student attrition rate, machine learning techniques such as classification can accurately predict students' dropout rate. This is an important task as identifying students at risk

of dropping out early can help prevent attrition among Government College students.

Classification has been successfully applied in various real-world domains and plays a crucial role in the education domain[9], particularly in predicting academic performance [3] – [8] . A review of 30 studies conducted between 2002 and early 2015 found that Artificial Neural Network (ANN), Decision Tree (DT), Naïve Bayes (NB), k-Nearest Neighbour (k-NN), and Support Vector Machine (SVM) were commonly used to build prediction models, with ANN and DT models yielding higher accuracy results. In recent years, there has been significant growth in research focused on predicting student performance, including predicting course drop-out/retention, predicting performance in Massive Online Open Courses (MOOCs) [14], and predicting students at risk of not graduating college on time [15], using classification (supervised learning) techniques[10-13].

The application of classification techniques in the education domain in Malaysia has mainly focused on student performance, with less attention given to attrition. However, a study by the author [16] aimed to identify the key factors influencing drop-out rates in Computer Science courses. The study collected demographic information and transcript records of students, focusing on core courses that had the most impact on drop-out cases. Four classification techniques, including k-NN, DT, NN, and Logistic Regression (LR), were employed to classify the dataset. The findings showed that the LR classifier was the most accurate, achieving 91% accuracy compared to other techniques used in the study. The results of the study indicate that there are five important courses in which students must score higher to reduce the risk of dropping out.

Bedregal Alpaca et al. [17] proposed a classification model that utilizes academic information from universities to identify students at risk of dropping out. The model considers various data such as demographic, academic performance, admission test, and course information for the evaluation. The result shows that the model can determine the most significant variable that affects academic performance, which is the number of abandoned subjects. In a similar vein, Gil, Delima, and Vilchez [18] applied DT and NB to identify the underlying factors of student drop-out in a public school in the Philippines. They utilized the Weka toolkit to employ the classifier algorithm on the selected dataset and produced a comparative performance result for each algorithm in terms of recall, precision, and accuracy.

Meanwhile, [19] focused solely on the k-NN technique to extensively evaluate and predict early-stage student drop-out. The technique is versatile, simple, and can handle various types of data. The results can help teachers identify students at risk of drop-out and monitor their well-being. Mardolkar and Kumaran [20] used data mining techniques to develop comprehensive

prediction models of student drop-out as early as possible. The models with sufficiently high accuracy will be used in an early warning system to detect students at high risk of drop-out as soon as possible. They explored academic variables (both at universities and former schools), sociodemographic factors, behaviors, and extracurricular activities that may influence student drop-out. However, only a subset of attributes that have a very high predictive contribution to student drop-out was considered.

Tomasevic, Gvozdenovic and Vranes [21] aimed to provide a detailed comparison of various supervised machine learning techniques to identify students who are at risk of dropping out from their course. They evaluated the performance of different classifiers, including k-NN, SVM, ANN, DT, NB, and LR. Their study found that ANN produced the highest precision by using past performance data and student engagement data in online learning.

Viloria and Padilla [22] applied NN, DT, and Bayesian Network to predict drop-out rate among engineering students in India. They found that academic results and socioeconomic situation have an impact on students, and managing these variables can help reduce drop-out rate. Similarly, Sangodiah et al. [23] used SVM to predict academic performance for students who were under probation in a private higher learning institution and obtained an 89.84% accuracy. Likewise, [24] used Binary Linear Regression to predict postgraduate doctoral degree students who would complete their study on time, and found that only 6.8% of the students in the year 2014 were able to graduate on time.

4. RESEARCH METHODOLOGY:

The aim of this study was to investigate the underlying reasons for student dropouts at the undergraduate level, with the following specific objectives:

1. To identify the primary academic, social, family, financial, job-related, and personal factors that contribute to dropout among students at the graduation level.
2. To examine the impact of student dropout on society.
3. To propose recommendations for reducing the rate of student dropouts.

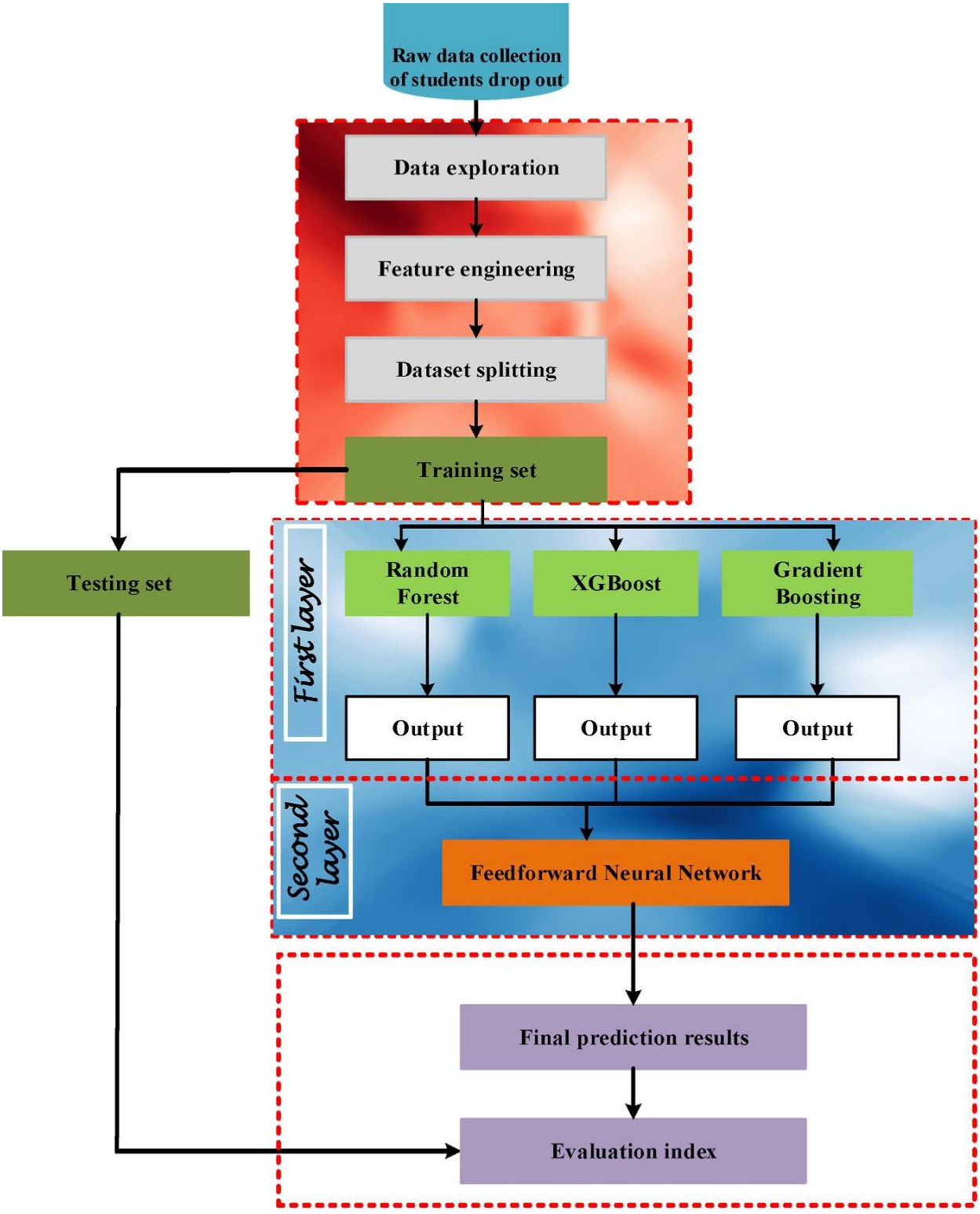

Figure 1: Our Proposed Methodology

The results of this research could be valuable for policymakers in developing strategies to address dropout rates in educational institutions, as well as for promoting quality education and economic progress. The study focused on graduate students at Government Arts and Science College in Arakkonam, Tamilnadu, where there are nine programs offered (B.Sc Computer Science, BCA, B.Sc Mathematics, B.Sc Zoology, B.Com General, B.Com CA, BBA, B.A Tamil, B.A English). At the time of the study, there were 13,272 students enrolled in the programs, with 326 students identified as dropouts. A sample of 31% of the total dropouts was selected using a simple random sampling technique.

The study aimed to fulfill its objectives by purposefully selecting a sample of 100 students who had dropped out. A structured questionnaire was used to collect data, which was obtained from classrooms and administrative officers involved in semester-wise course registration. Information such as the names, roll numbers, and contact details of the dropout students were collected, and they were interviewed over the phone. The university's student information server was used to facilitate contact with the dropout students. The collected data was tabulated and analyzed using both the Statistical Package for the Social Sciences (SPSS) and Microsoft Excel.

5. EXPERIMENTAL RESULTS AND DISCUSSION:

The proposed approach in this study was primarily implemented in Python, using scikit-learn (0.24.2), a popular package developed by Google (Yang et al., 2021). The study proposed a stacking ensemble, based on a hybrid of RF, XGBoost, GB, and FNN, to accurately predict student dropouts in our Government College. Currently, only Random Forest is used for the stacking ensemble. A cross-validation approach was adopted to avoid over-fitting, and the input data were randomly divided into training and testing datasets with 80% and 20% percentages, respectively. Ten-fold cross-validation was applied to address over-fitting issues. The performance metrics discussion confirms that the selected binary classification models can predict students' dropout or non-dropout at the individual course level, even with a limited number of input features. However, before using the classifiers to predict student dropouts, a broader set of performance metrics needs to be investigated, which is consistent with findings in other research papers in the domains of AI and ML.

The case study and its findings have some limitations, as mentioned earlier. The primary limitation is the limited size of the dataset. Unlike other ML model application domains, it is difficult to increase the amount of data in the educational domain by combining different resources. This is because individual records often represent the learning outcomes or behavior

of individual students, resulting in datasets that are smaller than what the ML algorithms require. To overcome this limitation, it is often necessary to employ a rigorous research design similar to that used in classical educational technology research, in which a sufficient number of records is estimated or guaranteed before the experiment begins.

The same applies to the quantity and quality of the independent features. In the learning process, the repetition of certain activities is often utilized. If these activities are interdependent or conditioned, their integration into the ML model should be carefully considered before beginning the data collection process. A weakness in the study was also identified in selecting the time threshold for making the prediction. As the time variable was not directly included, it was necessary to identify specific points when the students' performance would be challenging to forecast.

The end of the second third of the term was used as a compromise, considering the natural distribution of individual activity categories in the course sections. However, more favorable outcomes could be anticipated if the course activities were designed with ML techniques in mind. Another limitation of the study is that different course runs generated diverse data to analyze, making it difficult to identify which attributes are generally important for predicting student performance as shown in Table 1.

Table 1: Simulation Parameters

Parameter	Description	Value/Details
Dataset	Student data from Government Arts and Science College	2016-2022
Sample Size	Number of dropout students surveyed	100 out of 326
Algorithms Used	Machine learning algorithms for prediction	Random Forest (RF), Feed-forward Neural Networks (FNN)
Model Type	Stacking ensemble model	Hybrid of RF and FNN
Data Split	Training and testing data split ratio	80% training, 20% testing
Cross-validation	Technique to prevent overfitting	10-fold cross-validation
Performance Metrics	Metrics for model evaluation	Testing accuracy, Area Under the Curve (AUC)
Data Collection Methods	Methods used to gather data	Surveys, interviews
Key Dropout Factors	Identified factors contributing to dropout	Study pressure, financial problems, desire to study abroad
Software and Tools	Implementation tools	Python, scikit-learn (0.24.2)
Data Analysis Tools	Tools for data analysis	SPSS, Microsoft Excel

Dropout is a widespread problem in the education sector not only in Tamilnadu but worldwide. Government Arts and Science College, Arakkonam, is one of the newest government

colleges, which commenced functioning on May 12, 2012. In the first batch, 323 students were admitted, and as of August 2022, there are a total of 13,272 registered students across the nine courses. Among them, 326 students have been identified as dropouts, representing 2.43% of the total student population.

- FILE DESCRIPTIONS:

The file descriptions for the dataset used in this study are as follows. The Student Academic Progress Data file contains detailed records of students' academic performance over time. The Student Static Data file includes static demographic and background information about the students. The Student Financial Aid Data file provides information on the financial aid received by the students. The Test Data file consists of student IDs for which dropout predictions need to be made. A Sample Submission File is provided, demonstrating the correct format for submissions. The Data Dictionary offers supplemental information about the dataset, explaining the various attributes and their meanings. The DropoutTrainLabels.csv file contains student IDs along with labels indicating whether each student dropped out or not, and is used for model building.

Our sample size was 100 out of the 326 dropout students. Based on sample survey, the following analysis in total is shown in the bar chart. It is seen in the chart that the reason of the highest dropouts relates to academic problem and the lowest dropout is insecurity. It can be guessed from the chart that the amount of learning that GASC takes from the students is very low. It is found from the analysis that the level of security at GASC is very good as only 0.5% student said that dropout was due to insecurity as shown in Table 2. The above scenario can also be displayed through a pie diagram. From the pie diagram, it is seen that financial crisis is the chief cause of dropout. The second highest reason mentioned by the dropout students is the reason for higher pressure of study. Besides, various reasons are related to dropout who includes job-related problem, family problem, going abroad, changed university, and security problem as shown in Table 2.

Table 2: Reasons for Drop out

DEGREE	**Reason**					
	Course	**Family**	**Finance**	**Study Pressure**	**Other**	**Total**
B.Sc Computer Science	1	5	6	1	2	15
BCA	3	8	10	1	2	24

B.Sc Mathematics	1	4	5	5	3	18
B.Sc Zoology	4	5	9	1	1	20
B.Com General	8	10	20	2	4	44
B.Com CA	8	10	24	3	3	48
BBA	7	10	20	1	2	40
B.A Tamil	8	15	26	0	2	51
B.A English	9	22	31	2	4	68

The faculty members at GASC are committed to providing quality education to their students. However, our analysis has revealed that guardians are highly concerned about their children's careers. When a student is struggling in their courses, the guardian becomes worried and may ultimately withdraw the student from the college. The study has shown that 20% of dropouts were due to academic pressure, 5% were due to weak English skills, 4% were unable to adjust to the environment, 8% failed courses, and 3% were dropped out because their credits from other colleges were not accepted. Students with weak English skills struggle to understand the course material, which makes them bored and leads to dropping out. A pie chart can also display the reasons for study-related dropouts. The Finance section covers a significant portion of the circle. Despite these challenges, the faculty members remain dedicated and serious about providing quality education to their students.

The college's affordable tuition and high-quality amenities have garnered praise from both guardians and education experts, resulting in its growing popularity. Students have compared the fees of this college to those of other institutions offering similar levels of education, and have found it to be a more cost-effective option. The reasons for dropout are numerous and differ by gender. For girls, marriage and domestic responsibilities, particularly childcare, as well as financial difficulties at home, are common factors leading to a high dropout rate. For boys, financial constraints, poor academic performance, and the desire to contribute to the family's income are the main reasons for dropping out.

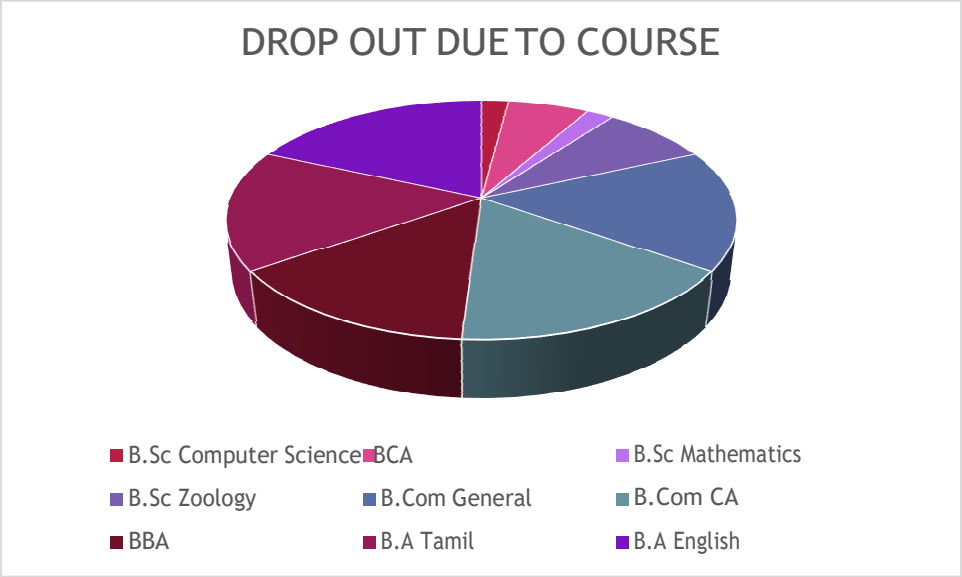

Figure 2. Drop Outs due to Course

Based on the data provided, it appears that there are various reasons for students dropping out of government colleges across different degree programs. Let's take a closer look at each degree program and the reasons provided for students dropping out:

B.Sc Computer Science: Among the 15 dropouts in this program, 6 cited financial reasons as their primary reason for dropping out, followed by 5 who cited family pressure as their reason. This could indicate that some students are struggling to balance the cost of education with their other financial responsibilities, while others may be facing pressure from their families to pursue a different path.

BCA: This program had the highest number of dropouts, with 24 students leaving for various reasons. Like B.Sc Computer Science, financial reasons were the most common, cited by 10 students. Family pressure was the second most common reason, cited by 8 students. Given that both B.Sc Computer Science and BCA are computer-related programs, it's possible that students in these fields are finding it difficult to keep up with the rapidly changing technological landscape, leading to increased pressure and stress.

B.Sc Mathematics: Of the 18 students who dropped out of this program, the most common reason given was a tie between financial pressure and study pressure, with 5 students each citing these reasons. This could indicate that students are finding the program to be academically challenging while also struggling to finance their education.

B.Sc Zoology: Among the 20 students who dropped out of this program, the most common reasons given were financial pressure (9 students) and family pressure (4 students). It's possible that students in this program are finding it difficult to balance the cost of education with other financial responsibilities, while also facing pressure from their families to pursue other

career paths.

B.Com General: This program had the highest number of dropouts among the commerce-related degrees, with 44 students leaving for various reasons. Financial pressure was the most common reason cited by students (20 students), followed by family pressure (10 students). This could indicate that students in this program are finding it difficult to balance the cost of education with other financial responsibilities, while also facing pressure from their families to pursue other career paths.

B.Com CA: Like B.Com General, financial pressure was the most common reason cited by students in this program who dropped out (24 students). Study pressure was also a common reason, cited by 3 students. This could indicate that students in this program are finding it difficult to keep up with the rigorous academic demands of the program while also struggling to finance their education.

BBA: Among the 40 students who dropped out of this program, the most common reason given was family pressure (10 students), followed by financial pressure (7 students). This could indicate that students in this program are facing pressure from their families to pursue other career paths, while also struggling to finance their education.

B.A Tamil: Among the 51 students who dropped out of this program, the most common reason given was financial pressure (26 students), followed by family pressure (15 students). This could indicate that students in this program are finding it difficult to balance the cost of education with other financial responsibilities, while also facing pressure from their families to pursue other career paths.

B.A English: This program had the highest number of dropouts overall, with 68 students leaving for various reasons. Like B.A Tamil, financial pressure was the most common reason cited by students (31 students), followed by family pressure (22 students). It's possible that students in this program are finding it difficult to balance the cost of education with other financial responsibilities, while also facing pressure from their families to pursue other career paths.

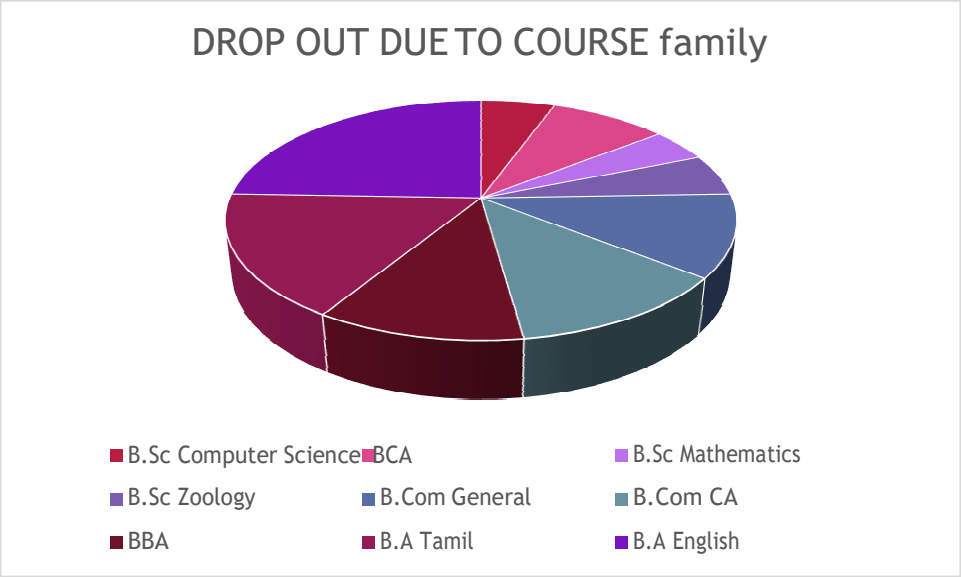

Figure 3: Family Factors affecting students drop out

In this work, it has been found that 70% of the students who dropped out due to financial crisis experienced the crisis after enrolling in the college. Another 10% dropped out because their waiver was cancelled. Financial crisis can result from various reasons such as sudden unemployment of the main earning member or a drop in crop prices for farmers. For instance, many students who cited financial crisis as the reason for dropping out mentioned that their fathers or brothers lost their jobs. At GASC, students are allowed to take the final exam even if they have outstanding fees for the whole semester. However, 20% dropped out due to accumulating a large amount of dues. The university was undergoing changes during this period.

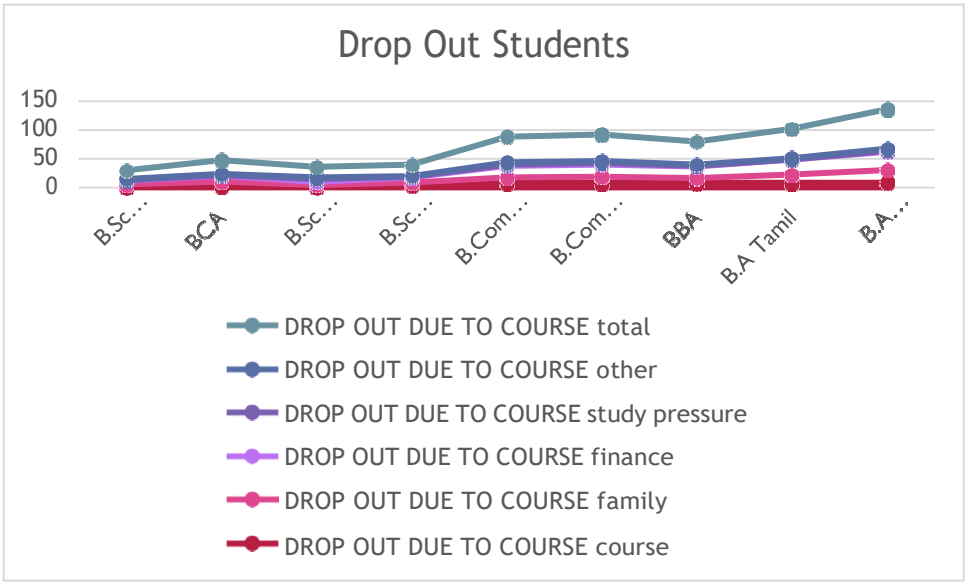

Figure 4: Total no. of Students Dropouts.

In Arakkonam, many women are unable to pursue education due to factors such as early marriage, security concerns, and financial constraints. Unfortunately, it has been observed that after marriage, many girls are discouraged or prevented from attending schools, colleges, and

universities. As a result, family-related issues such as early marriage, the inability to bear education expenses, and security problems are major causes of female students dropping out. Parents sometimes force their daughters to leave school or university, and in some cases, husbands and other guardians prohibit women from continuing their education. Out of 18 female dropouts, 14 (77%) left due to family-related issues, including separation of parents, death of a family member, and accidents involving family members.

6. RECOMMENDATION:

To control the worse situation of students' dropouts from the GASC, the authority has planned to take some remedial measures such as:

1. To motivate students, the faculty members should increase their counseling hours and take on additional responsibilities.
2. Remedial English courses should be introduced, and only those who pass the course should be allowed to enter the mainstream programs.
3. A trustee could be established to help support bright and financially disadvantaged students.
4. Campus jobs could be offered to outstanding yet impoverished students.
5. Donations could be sought from various donor agencies to provide assistance to needy students.
6. Study loans could be arranged through both banks and other NGOs.
7. Study loans could be provided to offer financial aid to students.
8. The Trustee could offer substantial scholarships for exceptional and underprivileged students.
9. While there may be limited options for addressing family-related problems, the college could hold community awareness seminars to promote the importance of education and discourage dropping out. The college could also launch a campaign to raise awareness and encourage families to allow girls to attend university.

7. IMPLEMENTATION:

The process of data preparation includes activities such as data cleaning, data joining, transforming attributes, and selecting or reducing features. To reveal hidden patterns in the data, different techniques can be employed based on the understanding of the data. The result of this stage is the creation of a final dataset that is considered suitable for input into the modeling tool. Educational data is typically sparse, noisy, inconsistent, or contains numerous attributes that are not relevant.

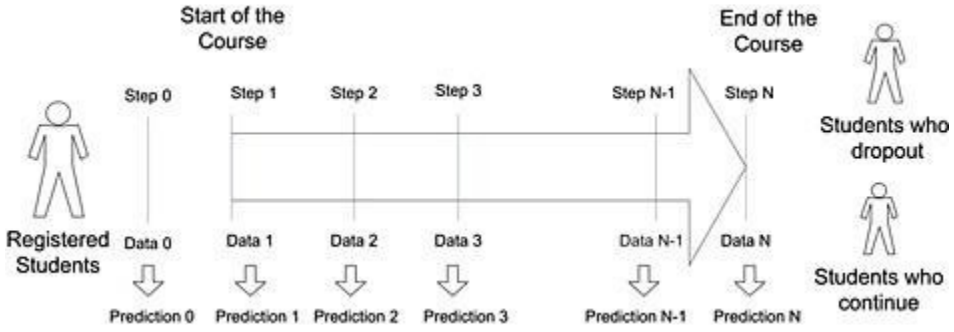

Figure 5: Prediction levels

The goal of the data preparation phase is to eliminate undesired data characteristics, such as noisy, inconsistent, or irrelevant attributes, and to handle missing values. To address noisy data, methods like binning, regression, or clustering can be used. Missing values can be dealt with by removing incomplete records, imputing them with global constants or the most probable value, but the semantics of missing values should be evaluated on a case-by-case basis. Python, with its Pandas and Numpy libraries for data manipulation, and Scikit-learn library for various supervised or unsupervised machine learning methods, was used for this project. Additionally, Keras and Tensorflow were utilized for neural network construction. Although there are low-code machine learning libraries available, their use in the LA domain is not yet widespread. Due to the scarcity of the educational dataset, the PyCaret library was employed to add several classifiers, and their performance was compared using randomly divided training and testing sets. The aim was to find a precise prediction point to identify students at risk of dropping out, which can be challenging with a small dataset. Confidence intervals were calculated for all performance metrics.

Accuracy is the overall accuracy rate or classification accuracy, and it is determined as follows:.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The dataset used in this project is included as studentdata.csv. This dataset has the following attributes:

- school ? student's college (binary: "GP" or "MS")
- sex ? student's sex (binary: "F" - female or "M" - male)
- age ? student's age (numeric: from 15 to 22)
- address ? student's home address type (binary: "U" - urban or "R" - rural)
- famsize ? family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
- Pstatus ? parent's cohabitation status (binary: "T" - living together or "A" - apart)
- Medu ? mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 -€“ 5th to

9th grade, 3 - secondary education or 4 -€“ higher education)

- Fedu ? father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 -€“ higher education)
- Mjob ? mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- Fjob ? father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- reason ? reason to choose this college (nominal: close to "home", collegel "reputation", "course" preference or "other")
- traveltime distance? home to school travel time (numeric: 1 - 1 hour)
- Teacher ? dropout (numeric: yes =1 no=0)
- Basic facility? college basic facility(numeric: yes=1, no=0)
- Family ? dropout for family background (numeric: yes=1, no=0)
- Health ? dropout for health problem (numeric: yes=1, no=0)
- Child Marriage ? dropout for child marriage (numeric: yes=1, no=0)
- Society? Dropout reason for society (numeric: yes=1, no=0)
- Economy ? dropout for family economy problem (numeric: yes=1, no=0)
- Disinterest ? not interest in education (numeric: yes=1, no=0)
- Psychology ? some psychological problem (numeric: yes=1, no=0)
- Parent compulsion ? dropout for Parent compulsion (numeric: yes=1, no=0)
- Unavailability Hr education ? nearest Unavailability Hr education (numeric: yes=1, no=0)
- Classmates ? problem for classmates (numeric: yes=1, no=0)
- Inability ? study for Inability (numeric: yes=1, no=0)
- Parents not realize education ? the parents are uneducated (numeric: yes=1, no=0)
- dropout ? did the student drop the class(character:(basic facility +family +economy

There are two stages in Random Forest algorithm, one is random forest creation, the other is to make a prediction from the random forest classifier created in the first stage. The whole process is shown below, and it’s easy to understand using the figure 5.

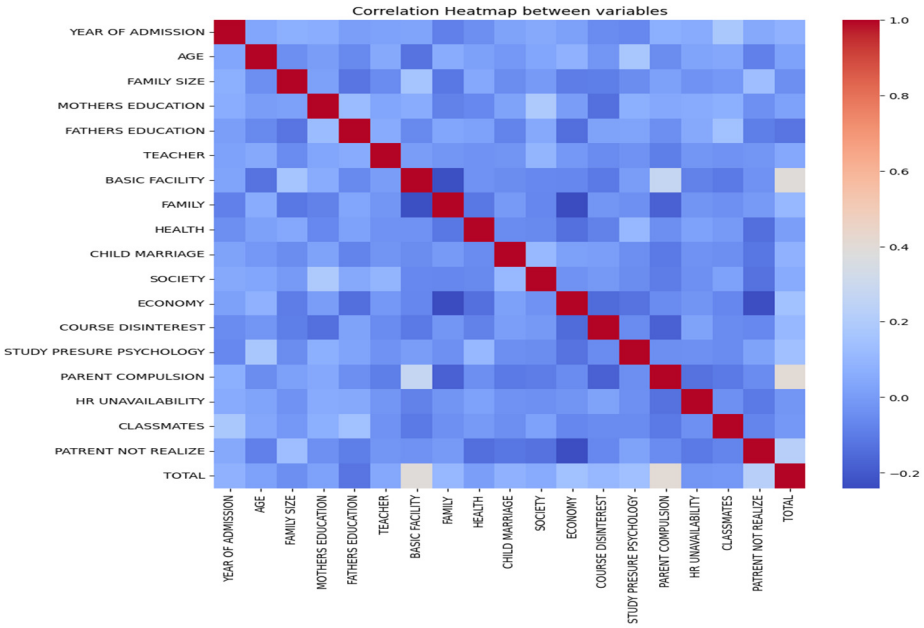

Figure 6.RF Prediction Levels

To assess the effectiveness of the proposed model, its experimental results were compared to the base models in the first layer, which included ANN, Decision Tree, RFM, XGBoost, and KNN. The comparison was based on training and testing results, as shown in Table 1. The training accuracy results ranged from 88.1% to 97.67%, while the testing accuracies ranged from 75.21% to 90.76%. It should be noted that the ANN suffered from overfitting issues, whereas the proposed method demonstrated improved prediction performance, achieving a training accuracy of 92.43% and a testing accuracy of 93.83%.

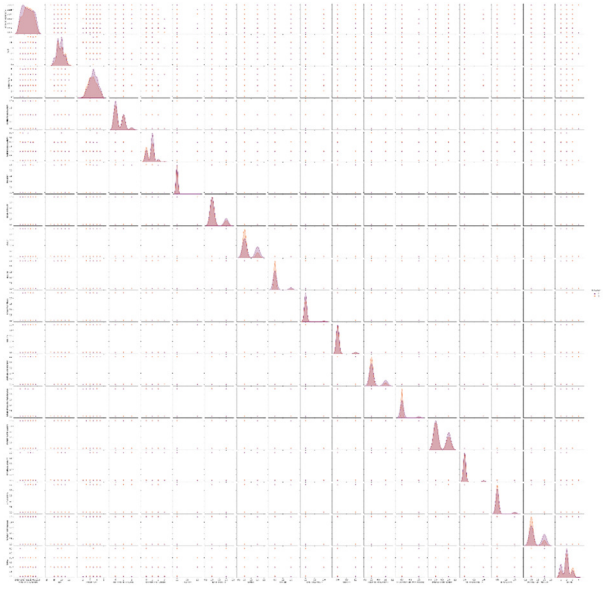

Figure 7 Heat map

Table 3 also showed that training the model was computationally demanding for all the models, with varying durations. The proposed RFM was the fastest and scientifically reduced the computational cost compared to the other techniques, with a runtime of 7.74 s. In contrast, the ANN's computation time was longer than the remaining models due to its hidden layers, which naturally recorded a significantly longer time. However, given the proposed problem's time frame, the runtime was not prohibitively long, taking around 12.49 s to run with all of the attributes enabled in the worst-case scenario.

Table 3: Computational Model

```
def comparemodel(x_train,x_test,y_train,y_test):
  print("Decision Tree Model")
  decisionTree(x_train,x_test,y_train,y_test)
  print('-'*100)
  print("Random Forest Model")
  RandomForest(x_train,x_test,y_train,y_test)
  print('-'*100)
  print("XGBoost Model")
  xgboost(x_train,x_test,y_train,y_test)
  print('-'*100)
  print("KNN Model")
  KNN(x_train,x_test,y_train,y_test)
  print('-'*100)
comparemodel(x_train,x_test,y_train,y_test)
```

In general, the recall results of 0.49 obtained by the stacking ensemble on the testing set suggest that the model can predict nearly 95% of dropout instances. Likewise, the precision value of 0.24 indicates a correctness of approximately 93% in predictions, covering both dropout and non-dropout instances among university students after course delivery. These results suggest that our method can efficiently handle every class, and ultimately, our model is better suited for predicting students' dropout in university classes.

Regarding the F1-Score, its values range from 0.66 to 0.32, where 0 indicates extremely poor results, and a value close to 1 demonstrates efficient results. Our method achieved an overall F1-Score of 0.92, indicating better prediction performance and making it a viable option for predicting students' dropout in college classes.

Confusiion matrix

[[112 0]

[117 0]]

Classification report

	precision	recall	f1-score	support
0	0.94	1.00	0.66	112
1	0.92	0.00	0.00	117
accuracy			0.49	229
macro avg	0.24	0.50	0.33	229
weighted avg	0.24	0.49	0.32	229

Figure 8: Prediction results

This classification report shows the performance of a binary classification model on a dataset with 229 instances. The report presents precision, recall, and F1-Score metrics for two classes, 0 and 1.

For class 0, the model achieved a high precision of 0.94, indicating that among the instances predicted as class 0, 94% were actually class 0. The recall score of 1.0 means that the model correctly identified all instances of class 0 out of the total of 112 instances in that class. The F1-Score of 0.66 shows a reasonable balance between precision and recall for class 0.

For class 1, the precision of 0.92 indicates that among the instances predicted as class 1, 92% were actually class 1. However, the recall score of 0.0 indicates that the model failed to identify any instances of class 1 out of the total of 117 instances in that class. The F1-Score of 0.0 suggests that the model's performance for class 1 is poor.

The overall accuracy of the model is 0.49, indicating that the model predicted correctly for less than half of the instances in the dataset. The macro-averaged F1-Score of 0.33 suggests that the model's performance is sub-optimal for both classes, while the weighted average F1-Score of 0.32 suggests that the model's performance is lower on the class 1 instances, which have a larger number of samples.

The analysis of different performance metrics affirms that the chosen binary classification models are suitable for forecasting whether students will drop out or not in specific

courses, despite having a small dataset with limited input features. Nevertheless, a wider range of performance metrics must be explored before utilizing the classifiers for dropout prediction. This conclusion aligns with the results reported in other research papers in the fields of AI and ML.

Despite the study's results, there are certain constraints to consider. Firstly, the dataset size is restricted, as mentioned earlier. Unlike in numerous other ML domains, combining various resources to enhance the volume of data in the educational field is not straightforward. This is due to the fact that individual records usually represent individual students' performance or conduct, resulting in datasets that are typically smaller than what the ML algorithms necessitate. This challenge is frequently overcome by adopting the same rigorous research design utilized in conventional educational technology research, which involves estimating or ensuring a sufficient number of records before the actual experiment is conducted.

The limitations extend to the number and quality of independent features. In the learning process, certain activities are often repeated, and their inclusion in the ML method should be carefully considered if they are interconnected or conditioned. Additionally, weakness was identified in selecting the time threshold for predicting student dropout. Because the time variable was not included, it was necessary to identify milestones when predicting student performance became difficult. The end of the second third of the term was chosen as a compromise based on the natural distribution of individual activity categories in course sections. However, designing course sections with ML techniques in mind would likely produce better results.

The difficulty in determining which attributes are important for predicting student performance is one of the limitations identified in the analysis. Another limitation is the choice of classifiers used in this case study. While there are more advanced ML techniques available for predicting early student dropout, none have produced significantly better results. The main goal of this study was not to find the best classifier, but it is worth noting that good predictions could be achieved using stacking ensemble, although performance metrics must be evaluated.

The study can be expanded in the future by utilizing regression techniques to forecast the occurrence of attrition by including data on students who are still enrolled and the precise date of drop-out. Additionally, the association rule approach can be employed to reveal concealed patterns that can be utilized to detect students who are at risk. The findings can then be authenticated by professionals from universities or the ministry.

8. CONCLUSIONS:

The issue of student dropout is a complex and negative problem in the education system, and it is a concern for various stakeholders, including parents, government institutions, and other educational agents. Despite efforts to address this problem, the consequences of student dropout remain unavoidable. However, accurate prediction of student dropout using analytical machine learning solutions can help reduce its social and economic costs. These solutions can effectively identify and predict influential factors such as social welfare, learning conditions, learner's outcome, age, gender, family status, and student sponsorship, among others, that contribute to student dropout. By predicting dropout accurately, students can focus on their studies and avoid the risk of dropping out, while teachers can intervene early in uncontrolled behavior that could lead to dropout risk and take proactive measures before the issue arises.

The primary goal of this study is to improve the performance of predicting college student dropout. The prediction can estimate the risk of students quitting their academic course and the resulting reduction in student outcomes. Furthermore, the study aims to develop an effective stacking ensemble-based method to minimize the negative impact of inaccurate student dropout prediction using a single model. The research aims to address the issue of student dropout at the course level. The results of the study demonstrate that, despite a small dataset, appropriately selected indicators that do not require access to system logs can be useful in predicting student dropout using different performance metrics. The predictive models used in this study were based on data gathered about students' learning environment activities and partial achievements. The outcomes obtained from the proposed method can aid in reducing the dropout rate by identifying the students who are likely to be affected and the influential factors contributing to dropout risk.

Overall, the data suggests that financial pressure is the most common reason for dropout among students in this college. Family pressure and course pressure are also major factors contributing to the high dropout rates. These findings may be useful in developing interventions to support students in these programs and prevent dropout.

In future research, it is suggested to explore other computational approaches such as deep learning and hybrid models for predicting student dropout and compare their results with those obtained in this study. Furthermore, other influential factors that were not considered in this study must also be examined, and a feature analysis study is recommended. Such studies can aid educational agents in effectively addressing the issue of student dropout.

9. REFERENCES

[1] Abu, R. Hamdan, R. and N.S. Sani, "Ensemble Learning for Multidimensional Poverty Classification," Sains Malaysiana., vol. 49(2), pp.447-459 2020.

[2] N. S. Sani, M. A. Rahman, A. A. Bakar, S. Sahran, and H. M. Sarim, "Machine learning approach for bottom 40 percent households (B40) poverty classification," IJASEIT, vol. 8, pp. 1698-1705, 2018.

[3] J. D. Holliday, N. Sani, and P. Willett, "Calculation of substructural analysis weights using a genetic algorithm," J. Chem. Inf. Model, vol. 55, pp. 214-221, 2015

[4] J. D. Holliday, N. Sani, and P. Willett, "Ligand-based virtual screening using a genetic algorithm with data fusion," Match-Commun. Math. Co., vol. 80, pp. 623-638, 2018.

[5] N. S. Sani, I. I. S. Shamsuddin, S. Sahran, A. H. A. Rahman, and E. N. Muzaffar, "Redefining selection of features and classification algorithms for room occupancy detection," IJASEIT, vol. 8, pp. 1486-1493, 2018.

[6] S. Shabudin, N. S. Sani, K. A. Z. Ariffin and M. Aliff, "Feature Selection for Phishing Website Classification," International Journal of Advanced Computer Science and Applications, vol. 11(4), pp. 587-595, 2020.

[7] T. K. M. Zali, N. S. Sani, A. H. Abd Rahman, and M. Aliff, "Attractiveness Analysis of Quiz Games," International Journal of Advanced Computer Science and Applications, vol. 10(8), pp. 205-210, 2019.

[8] Z. A. Othman, A. A. Bakar, N. S. Sani, and J. Sallim, "Household Overspending Model Amongst B40, M40 and T20 using Classification Algorithm," International Journal of Advanced Computer Science and Applications, vo1. 11(7), pp. 392-399, 2019.

[9] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," Procedia Comput. Sci., vol. 72, pp. 414–422, January 2015.

[10] M. A. Al-Barrak and M. Al-Razgan, "Predicting students final GPA using decision trees: A case study," Int. J. Inf. Educ. Technol., vol. 6, pp. 528–533, July 2016.

[11] E. A. Amrieh, T. Hamtini and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," 2015 IEEE Conf. Appl. Electr. Eng. Comput.Technol. (AEECT), Amman, Jordan, pp. 1–5. November 2015.

[12] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177–194, October 2017.

[13] F. Widyahastuti and V. U. Tjhin, "Predicting students performance in final examination using linear regression and multilayer perceptron," 2017 10th Int. Conf. Human Syst. Interact. (HSI), pp.

188–192, July 2017.

[14] R. Umer, T. Susnjak, A. Mathrani, and S. Suriadi, "Predicting student's academic performance in a MOOC environment," 11th Int. Conf. Data Mining, Comput., Commun. Ind. Appl. (DMCCIA-2017), pp. 119–124, December 2017. [Umer, R., Science, M., Susnjak, T., Mathrani, A., Science, M., & Suriadi, S].

[15] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. Addison, "Who, whenand why: A machine learning approach to prioritizing students at risk of not graduating high school on time," Proc. 5th Int. Conf. Learning Anal. Knowl., New York, pp. 93–102, March 2015.

[16] W. W. Yaacob, N. M. Sobri, S. M. Nasir, N. D. Norshahidi, and W. W. Husin, "Predicting student drop-out in higher institution using data mining techniques," J. Physics: Conf. Series 2020, vol. 1496, 012005, March 2020.

[17] N. Bedregal-Alpaca, V. Cornejo-Aparicio, J. Zárate-Valderrama, and P. Yanque-Churo, "Classification models for determining types of academic risk and predicting drop-out in university students, Int. J. Adv. Comput. Sci. Appl., vol. 11, pp. 266–272, 2020.

[18] J. S. Gil, A. J. P. Delima, and R. N. Vilchez, "Predicting students' dropout indicators in public school using data mining approaches." Int. J. Adv. Trends in Computer Sci. Eng., vol. 9, pp. 774–778, 2020.

[19] M. Mardolkar and N. Kumaran, "Forecasting and avoiding student dropout using the K-nearest neighbor approach," SN Computer Sci., vol. 1, pp. 1–8, March 2020.

[20] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S.Ventura, "Early drop-out prediction using data mining: A case study with high school students," Expert Systems, vol. 33, pp. 107–124. February 2016.

[21] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," Comput. Educ., vol. 143, p. 103676, January 2020.

[22] A. Viloria, J. G. Padilla, C. Vargas-Mercado, H. Hernández-Palma, N. O. Llinas, and M. A. David, "Integration of data technology for analyzing university drop-out," Procedia Comput. Sci., vol. 155, pp. 569–574, January 2019.

[23] A. Sangodiah, P. Beleya, M. Muniandy, L. E. Heng, and C. Ramendran, "Minimizing student attrition in higher learning institutions in Malaysia using support vector machine," J. Theoritical Appl. Inf. Technol., vol. 71, pp. 377–385, January 2015.

[24] S. S. Shariff, N. A. Rodzi, K. A. Rahman, S. M. Zahari, and S. M. Deni, "Predicting the "graduate on time (GOT)" of PhD students using binary logistics regression model," AIP Conf. Proc. 2016, vol.

1782, p. 050015, October 2016.

[25] K. Limsathitwong, K. Tiwatthanont, and T. Yatsungnoen, "Drop-out prediction system to reduce discontinue study rate of information technology students," Proc. 2018 5th Int. Conf. Business and Industrial Research: Smart Technol. Next Generation of Information, Eng., Business and Social Sci. (ICBIR 2018), pp. 110-114, May 2018.

[26] Y. Chen, A. Johri, and H. Rangwala, "Running out of STEM: A comparative study across STEM majors of college students at-risk of dropping out early," Proc. 8th Int. Conf. Learn. Anal. Knowl., pp. 270- 279, March 2018.

[27] J. M. Ortiz-Lozano, A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa, "University student retention: Best time and data to identify undergraduate students at risk of drop-out," Innovations Educ. Teach. Int., vol. 57, 74-85, January 2020.

[28] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student drop-out in higher education," arXiv preprint arXiv:1606.06364., June 2016.

[29] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Drop-out prediction in edx MOOCs," Proc. - 2016 IEEE 2nd Int. Conf. Multimedia Big Data, pp. 440-443, April. 2016.

[30] D. D. Pokrajac, K. R. Sudler, P. Y. Edamatsu, and T. Hardee, "Prediction of retention at historically black college/university using artificial neural networks," 2016 13th Symp. Neural Networks and Applications (NEUREL), pp. 1-6, November 2016.

[31] G. S. Abu-Oda, and A. M. El-Halees, "Data mining in higher education: University student drop-out case study," Int. J. Data Mining & Knowl. Manage. Proc., vol. 5(1), pp. 15-27, January 2015.

[32] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, and R. Abreu, "A comparative study of classification and regression algorithms for modelling students' academic performance," Proc. 8th Int. Conf. Educ. Data Mining, 392-395, January 2015.

[33] Siri, "Predicting drop-out at university using Artificial Neural Networks," Italian J. Soc. Educ., vol. 7, pp. 225-247, June 2015.

[34] Y. Chen, A. Johri, and H. Rangwala, "Running out of STEM: A comparative study across STEM majors of college students at-risk of dropping out early," Proc. 8th Int. Conf. Learn. Anal. Knowl., pp. 270– 279, March 2018.

[35] J. M. Ortiz-Lozano, A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa, "University student retention: Best time and data to identify undergraduate students at risk of drop-out," Innovations Educ. Teach. Int., vol. 57, 74–85, January 2020.

[36] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student drop-out in higher education," arXiv preprint arXiv:1606.06364., June 2016.

[37] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Drop-out prediction in edx MOOCs," Proc. - 2016 IEEE 2nd Int. Conf. Multimedia Big Data, pp. 440–443, April. 2016.

[38] D. D. Pokrajac, K. R. Sudler, P. Y. Edamatsu, and T. Hardee, "Prediction of retention at historically black college/university using artificial neural networks," 2016 13th Symp. Neural Networks and Applications (NEUREL), pp. 1–6, November 2016.

[39] G. S. Abu-Oda, and A. M. El-Halees, "Data mining in higher education: University student drop-out case study," Int. J. Data Mining & Knowl. Manage. Proc., vol. 5(1), pp. 15–27, January 2015.

[40] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, and R. Abreu, "A comparative study of classification and regression algorithms for modelling students' academic performance," Proc. 8th Int. Conf. Educ. Data Mining, 392–395, January 2015.

[41] A. Siri, "Predicting drop-out at university using Artificial Neural Networks," Italian J. Soc. Educ., vol. 7, pp. 225–247, June 2015.

. **Biographical notes**

S.Selvakani Working as an Assistant Professor and Head in the Department of Computer Science / Computer Applications at Government Arts and Science College, Arakkonam (Formerly Thiruvalluvar University College of Arts and Science). She Completed Ph.D (Computer Science), M.Tech (Computer Science and Engg), M.Phil (Computer Science), MCA (university III Rank) at Manonmanium Sundaranar University, Tirunelveli, Tamilnadu, India. She is a dynamic hardworking professional with rich experience of 22 years in Teaching, Administration/ Training & Development. She is having hands on experience in general administrative activities, Research contributions and Student Counseling. Cambridge University has recognized and awarded certification of "Teachers and Trainers" for remarkable achievement in the field of Teaching. She has published more than 80 papers in the peer reviewed International Journals and more than 6 papers in National Journals. Besides presented more than 112 papers in the National and International Conferences and guiding several Ph.D Scholars in Computer Science presented more than 112 papers in National and International Journals, Conference and Symposiums. She has published 9 books and 10 Patents. Her interest areas are Data Science, Big Data, Machine Learning and Network Security.

Mrs.K VASUMATHI Working as an Assistant Professor Department of Computer Science / Computer Applications at Government Arts and Science College, Arakkonam (Formerly Thiruvalluvar University College of Arts and Science). She Completed S.E.T (Computer Science), M.Phil (Computer Science), M.Sc (university VI Rank) at Thiruvalluvar University, Vellore, Tamilnadu, India. She is a dynamic hardworking professional with rich experience of 13 years in Teaching. She is having hands on experience in Research contributions and Student Counseling. She has published more than 18 papers in the peer reviewed International Journals. Besides presented more than 35 papers in the National and International Conferences. She has published 2 books. Her interest areas are Operating Sysyem, Big Data, Machine Learning.