

## Student Performance Prediction Using Educational Data Mining

<sup>1</sup>Prof. P. S. Ingle, <sup>2</sup>Anand Laxminarayan Jaiswal, <sup>3</sup>Raj Kailas Gore, <sup>4</sup>Mohammad Sofiyan Md. Aijaz, <sup>5</sup>Akshali Arun Parihar, <sup>6</sup>Kajal Arun Jadhao

<sup>1</sup>Department of Information Technology, Anuradha College of Engineering and Technology, Chikhli, India

<sup>2</sup>Department of Information Technology, Anuradha College of Engineering and Technology, Chikhli, India

<sup>3</sup>Department of Information Technology, Anuradha College of Engineering and Technology, Chikhli, India

<sup>4</sup>Department of Information Technology, Anuradha College of Engineering and Technology, Chikhli, India

<sup>5</sup>Department of Information Technology, Anuradha College of Engineering and Technology, Chikhli, India

<sup>6</sup>Department of Information Technology, Anuradha College of Engineering and Technology, Chikhli, India

\*\*\*

**Abstract - With the help of artificial intelligence (AI), education is going through a tremendous change. Through this change, it is now easier to get a better picture of how well students are doing and to spot kids who might quit school. This study looks at how learning analytics powered by artificial intelligence could be used to figure out how many students succeed and how many drop out by looking at patterns in students' attendance, online activity, and level of interest. The paper will use data from massive open online courses (MOOCs), online learning tools, and university databases to show how artificial intelligence can help students stay on track by allowing for timely interventions. It will be possible to do this by looking at the facts. That being said, it's possible that these treatments will include suggestions for adaptive learning and personalized feedback. The goal is to help teachers make smart choices, support students in doing well in school, and lower the number of students who drop out. This will be done with the help of an artificial intelligence model that can be trusted and is moral. To make sure that AI is used in schools in an acceptable way, this study will also look at problems like keeping data safe, getting rid of bias, and encouraging openness. At the end of the day, our study aims to make education more flexible, effective, and open to all students.**

*Keywords: student performance prediction, dropout rate analysis, machine learning in education, predictive analytics, adaptive learning, student retention, early warning systems, personalized education, data-driven decision-making.*

### I. INTRODUCTION

Educational Data Mining (EDM) is a fast developing subject that pulls from a range of disciplines, including data mining, machine learning, statistics, and visualization, in order to study the vast volumes of data that are generated inside educational environments. As one of the many essential uses of EDM, one of the most important applications is the capacity to forecast the amount of academic performance that students will achieve. In addition to these applications, there are a great many more that are of great significance. The purpose of this program is to identify children who are at risk of underperforming and to give timely interventions to those children who have been recognized as being at risk. All of these students will be included in the program. The utilization of this program allows for the identification of these youngsters to be completed. It is possible for educators and educational institutions to gain useful insights into the factors that drive academic achievement and to proactively take steps to increase learning outcomes through the application of a variety of machine learning algorithms. This is a capability that is available in the field of education. They have the ability to acquire these insights, which enables them to carry out this activity. Through the utilization of machine learning, this objective can be realized in a fashion that is both more expedient and less difficult.

In the process of seeking to ascertain the amount of academic performance that students have attained, it is a standard practice to take into account a number of major milestones that have occurred along the way. This is because determining academic performance is a process that involves a number of steps, which is the reason for this reason. The first phase in the process entails acquiring important information from a number of various educational platforms, including Learning Management Systems (LMS), student information systems, and

assessment databases, among others. Next, this information is brought into the system through the process of importation. The data in question could potentially contain a large variety of various properties, depending on the circumstances. There is a huge selection of choices available for those who are interested. The criteria that are included in this category include, but are not limited to, demographics, previous academic records, engagement indicators (such as the number of times a student logs into the learning management system or participates in online forums), assessment results, and even psychological or socio-economic characteristics. These are just some of the criteria that are included. After the data has been collected and preprocessed (which includes operations such as cleaning, addressing missing values, scaling features, and selecting features), the application of machine learning techniques is then carried out in order to construct prediction models. This is done in order to make predictions. As a result of this, it is possible to make

forecasts regarding subsequent events. As a consequence of this, it is feasible to make predictions concerning the events that will take place in the future. There is a wide variety of algorithms that can be utilized in order to achieve the most optimal combination of outcomes while adhering to the limits of the parameters that are available. This depends not only on the qualities of the data, but also on the purpose of the job that is being carried out, which is to make certain predictions. This is necessary in order to accomplish the task at hand. It is common practice to employ methods such as Logistic Regression, Support Vector Machines (SVM), and Decision Trees when dealing with tasks that involve binary classification. This is because these methods are able to accurately predict the outcomes of the work. Additionally, strategies that are analogous to ensemble approaches, like as Random Forests and Gradient Boosting, are widely utilized for projects of this sort. These strategies are used to complete assignments. Determining whether or not a student would be successful in a certain class is a good example of the kind of work that goes under this category. This is an example of the kind of work that falls under this category. A student will either pass the course or they will not pass it. There is no other option. There is a large number of regression models that have the potential to be effective when it comes to the practice of forecasting continuous outcomes, such as final grades. These models are available. Regression modeling is an example of a practice that falls into this category. Models such as linear regression, polynomial regression, and support vector

regression are examples of the types of statistical procedures that fall under this category. Furthermore, support vector regression is an additional example.

It is common practice to carry out tests and evaluations by making use of a variety of various performance indicators in order to choose the algorithm that is most appropriate for a certain set of circumstances. In order to ascertain the algorithm that will be chosen, this step is carried out initially. It is carried out in this manner in order to determine which algorithm is the most suitable for the situation that is being considered. When doing categorization assignments, it is usual practice to engage in the utilization of a number of different measurements. For example, accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC) are all examples of metrics that fall under this category. Several different measures are applied in order to fulfill the objective of providing assistance in the classification process. There are a number of measures that are utilized in order to assess the predictive capacity of the model. These metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. For the purpose of determining how predictable the model is, these metrics are utilized. At the time that regression operations are being carried out, these measures are utilized with the intention of determining whether or not the model is accurate. Because of the specifics of the situation, this is done. Using cross-validation techniques is crucial in order to prevent overfitting and to ensure that the model can be used with data that has not been seen before. This is because overfitting can be detrimental to the accuracy of the model. In order to accomplish this objective, it is necessary to make certain that the model can be applied with both types of data being taken into consideration. Because of the insights that can be derived from these prediction models, it is feasible that educational institutions and organizations will be able to make a substantial amount of progress. This is because of the potential for these models to provide. It will be possible to put into action treatments that are especially customized to suit the requirements of the children if it is proved that children who are at risk have been identified at an early level. This will be the case if it is demonstrated that the identification of these children has occurred at an early stage. therapies such as personalized feedback, one-of-a-kind learning tools, and further training are all examples of therapies that could be implemented within the context of the learning environment. There are also further instances, such as the availability of tools that allow for the student's learning to be individualized. It is

possible to achieve the objective of acquiring an understanding of the elements that greatly influence academic performance by going through the process of gaining an understanding of educational methods and the development of curriculum. This objective can be accomplished by going through the process of acquiring this understanding. It is conceivable for teachers to place an emphasis on and provide rewards for students who

actively participate in online talks if the model suggests a substantial association between active involvement in online chats and higher results, for example. This would allow teachers to present students with opportunities to be more involved in online conversations. Among these possibilities is the potential that the model will discover a strong connection between the two objects. There is a potential that could be utilized in this situation. However, it is of the utmost importance to acknowledge the ethical considerations that are involved with the use of machine learning for the purpose of predicting the level of success that students will face in their academic endeavors. This is because the goal of this application is to forecast the level of success that students will encounter. The rationale behind this is that the objective of this application is to provide an estimation of the level of success that students will achieve in their academic endeavors. For the purpose of addressing the issues regarding data privacy, algorithmic bias, and the possibility of discriminatory effects, a significant amount of caution is required. The resolution of these problems is of the utmost importance. The resolution of these concerns as quickly as feasible is of the utmost importance and should be done as soon as possible. When it comes to the process of designing and putting these models into action on the ground, it is of the utmost significance to make certain that justice, openness, and accountability are maintained throughout the entire process. The significance of this cannot be overstated. There is a considerable amount of potential for enhancing educational outcomes and forecasting the academic performance of students through the employment of machine learning algorithms in Educational Data Mining. This potential may be realized through the exploitation of Data Mining. There is a considerable amount of excitement that comes from this potential. This potential may be developed through the usage of educational data mining, which would be the mechanism by which it would be achieved. This potential comprises a large variety of diverse possibilities, which means that it offers a wide range of options and chances. Because of the capacity of data analysis, educators are able to discover

insights that can be put into action, which in turn enables them to provide students with support that is more effective. Everyone that is involved is going to come out ahead in this situation. It is of the utmost importance to give ethical concerns a higher priority and to make certain that these technologies are controlled in the appropriate manner in order to create a learning environment that is more equal and beneficial. This is because giving ethical concerns a greater priority is very important. This responsibility includes the management of the environment, which must be carried out in the appropriate manner in order to fulfill its obligations. The reason for this is because the region is consistently making progress, which is the reason why this is the case. This is the reason why this thing is happening.

II. LITERATURE REVIEW

Ref	Year	Author's	Key Technology	Results	Comment/Gap
1	2024	Dalia Khairy, Nouf Alharbi, Mohamed A. A. Mashayekha, Marwan F. Alsaif, Saleem Alkhalf, Rania A. Abougallala	Random Forest and Decision Tree Classification	The main goal of this study is to predict the performance of undergraduate first-level students in the Computer Department during the years 2016 to 2021 to enhance their performance in future by discovering the best algorithm	This was determined based on the findings of the study. Out of a total of 253 instances that were included in the testing set, they only made three incorrect classifications.

			use to analyze the educational data to identify the students' academic performance.	
2	20 24	M. A. Arulmiyan, Rozario, Dr. R. Guna Sundari	This paper presents a framework for predicting the academic performance of first-year bachelor's students in computer science courses using data mining techniques.	This research contributes to the field of educational data mining, offering insights that can potentially enhance student outcomes in computer science education.
3	20 24	Mohamed Bellajne, Ahmed Ben Dahan, Said Boudra, Mohammed Lama rti	This study aims to comprehensively evaluate the methods, tools, and applications of machine learning (ML) and data mining (DM) in education.	The ML models outperform other ensemble approaches, providing a valuable tool for predicting student performance and assisting educators in making proactive decisions through timely alerts.

		Sefi an		
4	20 24	Esma el Ahmed	Modern learning institutions face challenges in analyzing performance, providing high-quality education, formulating strategies for evaluating students' performance, and identifying future needs.	The outcomes of parameter adjustment greatly increased the accuracy of the four prediction models. Naïve Bayes model's prediction accuracy is the lowest when compared to other prediction methods, as it assumes a strong independent relationship between features.
5	20 25	Athanasios Angelopoulos, John Aliprantis, Markos Konstantakis, Alkiviadis Tsimpiris	This study investigates the use of educational data mining (EDM) techniques to predict student performance and enhance learning outcomes in higher education.	This study offers insights into the effective application of data-driven approaches to improve educational outcomes and foster student success.

6	20	Meh	Ed	Smart	Improvement
	24	mood	ge	City of the	technology integration
		a, H.,	Co	post-	drastically is required
		Khali	mp	cloud era	
		d, A.,	utin	is often	
		Kosta	g	envisione	
		kos,	d	as a	
		P.,		world in	
		Gilm		which	
		an,		virtualise	
		E., &		d	
		Pirti		resources	
		kanga		deliver	
		s. S		on-	
				demand	
				computin	
				g power	
				and	
				services.	

### III. RESEARCH GAPS

There have been a number of studies that have been carried out over the course of a number of years with the purpose of predicting the level of academic and behavioral performance of children based on the data that was gathered from those pupils. These studies have been carried out during the course of a number of years. The research for these studies has been conducted over the course of a number of years. In order to make a prediction about the level of intellectual and behavioral achievement that children have acquired, these investigations were carried out with the purpose of making a prediction. What evaluation methods are now in place are not particularly helpful when it comes to identifying children who are experiencing issues at an early stage. This is because these systems are not really beneficial. The reason for this is specifically due to the fact that the evaluation systems are not particularly successful. It is because of this that this is the case. It is vital to construct a machine learning model that is capable of properly forecasting the performance of students in order to achieve the aim of enabling timely intervention. This is because the goal is to enable timely intervention. The accomplishment of this objective is built upon this foundation, which is the source of the foundation. There is no possible way to overestimate the

significance of this particular stage of the process. It is important to fulfill a set of requirements in order to qualify for the utilization of this particular form of model. There are a number of models that have been investigated in order to explore learning patterns. Two examples of these models are convolutional neural networks and recurrent neural networks, which are both instances of deep learning models.

### IV. RESEARCH OBJECTIVES

- To design and implement a machine learning-based predictive model for forecasting students' academic performance with high accuracy, enabling early identification of at-risk learners.
- To develop a classification framework using deep learning algorithms to identify slow learners and fast learners, with the aim of supporting individual learner's capability-based teaching practices.
- To design and implement a machine learning based predictive model for suggesting suitable career choice to the students.

### V. RESEARCH METHODOLOGY

Collecting the data with the prescribed format so that all points related to objective will be covered, also we generate the link for collection of same data with the category of Graduation students

#### A. Data Collection and Preprocessing:

- Academic Records: Historical grades (specific subjects, cumulative GPA), attendance data, standardized test results (if applicable), course completion statistics, disciplinary history.
- Demographic Data: Age, gender, socioeconomic situation (anonymized), familial background (e.g., parents educational attainment).
- Behavioral Data: Involvement in extracurricular activities, interaction with online learning systems (e.g., login frequency, duration of participation, patterns of assignment submissions).
- Psychometric Data (Optional but Advisable): Self-efficacy metrics, motivational indices, learning modalities (by surveys or questionnaires).

*B. Data Preparation:*

- Addressing Missing Values: Employing imputation techniques (mean, median, mode, K-nearest neighbors imputation) or eliminating rows/columns with substantial missing data.
- Categorical Encoding: One-hot encoding for nominal variables (e.g., gender, course type) and ordinal encoding for ordinal variables (e.g., education levels).
- Outlier Identification and Management: Detect and address outliers utilizing methods such as Z-score, Interquartile Range (IQR), or robust scaling.
- Data Normalization/Standardization: Adjust numerical features to a uniform range (e.g., Min-Max scaling or StandardScaler) to ensure that features with greater magnitudes do not overshadow the model.

*C. Deployment and Surveillance of Models:*

- Batch Prediction: Consistently execute predictions on incoming student data.
- Real-time Prediction: Incorporate the model into an internet platform for immediate forecasts as fresh data is received.
- Detailed Evaluation Metrics: Scores from individual assignments, quizzes, and tests (both formative and summative); scores on specific topics or learning objectives within a subject; project evaluations, laboratory reports, and presentations; standardized test scores (if accessible); pre-assessment scores for particular units or topics; grades from prior courses, semesters, or academic years (e.g., GPA); number of failed courses or attempts at specific courses; and academic progress over time (e.g., fluctuations in grades).

- Temporal Performance Metrics: Quantity of attempts required to achieve mastery



of a concept or successfully complete an evaluation.

Fig. 1 Flow of the proposed system

VI. RESULT AND DISCUSSION

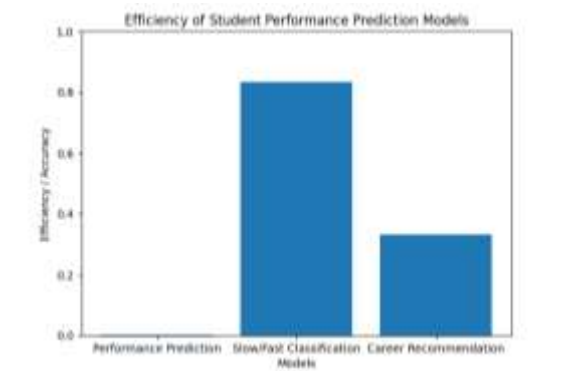


Fig. 2 Efficiency of Student performance

It provides a comparison of the three distinct machine learning models that have been constructed in terms of their effectiveness. This comparison can be seen in Figure 2, which can be found here. With an accuracy of approximately 83%, the slow and rapid learner classification model got the highest level of accuracy, indicating that it did remarkably well in recognizing different types of learners through its categorization. Because of the difficulty of making predictions across several classes and the limited size of the dataset, the career recommendation model was able to achieve a level of accuracy that was deemed to be moderate. It was discovered that the model for predicting academic accomplishment was not particularly effective, which brought to light the necessity of improving feature selection and expanding the amount of training data that was available. Evidence that the proposed framework is effective in learner classification is provided by the findings when they are considered as a whole. Furthermore, there is opportunity for further improvement in terms of performance prediction and career advising.

## VII. CONCLUSION

Machine learning has been shown to be effective in the early detection of kids who are at risk, as demonstrated by the academic performance prediction model that has been suggested. This makes it possible for educational institutions to move away from reactive measures and toward proactive treatments, which, in turn, leads to a reduction in the number of students who drop out of school and brings about an improvement in the overall academic outcomes. For the purpose of enhancing capability-based and individualized education, the classification of slow and fast learners, which is based on deep learning, provides unique insights that can be utilized. Every student, regardless of the rate at which they are learning, benefits from this since it helps to improve their engagement, knowledge, and performance. In addition, the career suggestion model indicates that the combination of academic, psychometric, and interest-based data can result in the development of accurate and tailored career advice being provided to the individual. Students are able to make judgments that are based on suitable information with the assistance of this, which significantly enhances workforce

alignment. The study, in its whole, lays the framework for educational approaches that are capable of being intelligent, equitable, and data-driven. These methods are

helpful to students, institutions, and the greater community as a whole.

## REFERENCES

- [1] Dalia Khairy, Nouf Alharbi, Mohamed A. Amasha, Marwa F. Areed, Salem Alkhalaf, Rania A. Abougalala. (2024). Prediction of student exam performance using data mining classification algorithms. *Education and Information Technologies*, 29:21621–21645
- [2] M. AArul Rozario, Dr. R. GunaSundari. (2024). Predicting and Analysis of Students' Academic Performance using Hybrid Techniques. *International Journal of Intelligent Systems and Applications in Engineering IJISAE*, 12(3), 4189–4195
- [3] Mohamed Bellaj, Ahmed Ben Dahmane, Said Boudra, Mohammed Lamarti Sefian. (2024). Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students' Academic Performance. *International Journal of Online and Biomedical Engineering (iJOE) iJOE*, Vol. 20 No. 3
- [4] Esmael Ahmed. (2024). Student Performance Prediction Using Machine Learning Algorithms. *Hindawi Applied Computational Intelligence and Soft Computing*, Volume, Article ID 4067721
- [5] Athanasios Angeioplastis, John Aliprantis, Markos Konstantakis, Alkiviadis Tsimpiris. (2025). Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach Using Educational Data Mining Techniques. *Computers*, 14, 83
- [6] Mehmooda, H., Khalid, A., Kostakos, P., Gilman, E., & Pirttikangas, S. (2024). A novel edge architecture and solution for detecting concept drift in smart environments. *Future Generation Computer Systems*, 150, 127–143.
- [7] Mahdia, O. A., Ali, N., Pardede, E., & Al-Quraishi, T. (2024). Online concept drift detector: Optimally balancing delay detection, runtime, memory, and accuracy. *Procedia Computer Science*, 237, 559–567.
- [8] Yawen Chen, Jiande Sun, Jinhui Wang, Liang Zhao, Xinmin Song, Linbo Zhai. (2025). Machine Learning-Driven Student Performance Prediction for Enhancing Tiered Instruction. *arXiv:2502.03143v1 [cs.LG]* 5 Feb.
- [9] Sazol Sarker, Mahit Kumar Paul, Sheikh Tasnimul Hasan Thasin, Md. Al Mehedi Hasan. (2024). Analyzing students' academic performance

- using educational data mining. *Computers and Education: Artificial Intelligence*, 7, 100263
- [10] Md. Mahmudul Islam, Farhad Hossain Sojib, Md. Fazle Hasan Mihad, Mahmudul Hasan, Mahfujur Rahman. (2025). The integration of explainable AI in Educational Data Mining for student academic performance prediction and support system. *Telematics and Informatics Reports*, Volume 18, June, 100203
- [11] Ahmed Kord, Ahmed Aboelfetouh, Samaa M. Shohieb. (2025). Academic course planning recommendation and students' performance prediction multimodal based on educational data mining techniques. *Journal of Computing in Higher Education*, Vol 2, 324
- [12] Venera Nakhipova, Yerzhan Kerimbekov, Zhanat Umarova, Laura Suleimenova, Saule Botayeva, Almira Ibashova, and Nurlybek Zhumatayev. (2024). Use of the Naive Bayes Classifier Algorithm in Machine Learning for Student Performance Prediction. *International Journal of Information and Education Technology*, Vol. 14, No. 1.
- [13] Arash Khosravi, Ahmad Azarnik. (2024). Leveraging Educational Data Mining: XGBoost and Random Forest for Predicting Student Achievement. *International Journal of Data Science and Advanced Analytics* (ISSN: 2563-4429).
- [14] Munaf Salim Najim Al-Din and Hussein Ali Al Abdulqader. (2024). *Students' Academic Performance Prediction Using Educational Data Mining and Machine Learning: A Systematic Review*. *International Journal of Research and Innovation in Social Science (Ijriiss)* ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume VIII Issue VIII August, Page 1264
- [15] Linqiang Tang, Chen Sian. (2025). Educational Data Mining For Student Performance Prediction. *Scalable Computing: Practice and Experience*, ISSN 1895-1767, SCPE. Volume 26, Issues 3, pp. 1551–1558
- [16] Abdullah Al-Ameri, Aniello Castiglione, Michele Nappi, Chiara Pero, Muhammad Umer. (2024). *Student Academic Success Prediction Using Learning Management Multimedia Data With Convolved Features and Ensemble Model*. *ACM J. Data Inform. Quality*, ACM 1936-1963/2024/8-ART
- [17] Bello, H. O., Ige, A. B., & Ameyaw, M. N. (2024). Adaptive machine learning models: Concepts for real-time financial fraud prevention in dynamic environments *World Journal of Advanced Engineering Technology and Sciences*, 12(2), 21–34.
- [18] Popoola, N. T. (2023). Big data-driven financial fraud detection and anomaly detection systems for regulatory compliance and market stability. *International Journal of Computer Applications Technology and Research*, 12(9), 32–46. <https://doi.org/10.7753/IJCATR1209.1004>
- [19] Saleem Malik, S. Gopal Krishna Patro, Chandrakanta Mahanty, Rashmi Hegde, Quadri Noorulhasan Naveed, Ayodele Lasisi, Abdulrajak Buradi, Addisu Frinjo Emma, Naoufel Kraiem. (2025). Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution. *Scientific Reports*, (2025) 15:8738
- [20] Evelina Staneviciene, Daina Gudoniene, Vytenis Punys, Arturas Kukstys. (2024). A Case Study on the Data Mining-Based Prediction of Students' Performance for Effective and Sustainable E-Learning. *Sustainability*, 16, 10442
- [21] Jedidi Y., Ibriz A., Benslimane M., Hachmoud A., Tmimi M., Hajjioui Y., Rahhali M. (2024). Predicting students' academic performance and modeling using data mining techniques. *Mathematical Modeling And Computing*, Vol. 11, No. 3, pp. 814–825
- [22] Esmael Ahmed. (2024). Student Performance Prediction Using Machine Learning Algorithms. *Applied Computational Intelligence and Soft Computing Volume*, Article ID 4067721, 15 pages
- [23] Murnawan, Sri Lestari, Rosalim Samihardjo, Deshinta Arrova Dewi. (2024). Sustainable Educational Data Mining Studies: Identifying Key Factors and Techniques for Predicting Student Academic Performance. *Journal of Applied Data Sciences*, Vol. 5, No. 3, September, pp. 1325-1342 ISSN 2723-6471 1325
- [24] Furkan AYDIN. (2024). Prediction of Academic Success and Selection of Effective Features with Machine Learning. *BUEFAD*, Volume 14, Issue 1, 14(1), 181-195
- [25] Şenay Kocakoyun Aydoğan, Turgut Pura, Fatih Bingül. (2024). Predicting Students' Academic Performances Using Machine Learning Algorithms in Educational Data Mining. *Malaysian Online Journal of Educational Technology*, (Volume 12 - Issue 4 )