ISSN: 2249-7196
IJMRR/Sep. 2024/ Volume 14/Issue 7/46-55
Mohd Shoaib / International Journal of Management Research & Review

A REVIEW ON DATA MINING AND MACHINE LEARNING METHODS FOR STUDENT SCHOLARSHIP PREDICTION

Mohd Shoaib[1], Omer Mohammed[2], Mohammed Raziuddin Faisa[3], Khutaija Abid[4]
[1,2,3] B.E. Student, Department of IT, Lords Institute of Engineering and Technology, Hyderabad
[4]Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad
khutaija@lords.ac.in

Abstract: *This review paper examines the application of Machine Learning and Data Mining techniques for predicting student scholarships. It conducts a comprehensive literature survey on the methodologies used in this area, emphasizing the significance of datasets in achieving accurate predictions. Machine Learning has gained considerable traction across various industries, including IT, education, and business sectors. The study discusses several ML/DM algorithms such as Naïve Bayes, Decision Tree, and k-NN, highlighting their effectiveness in determining scholarship eligibility. The proposed model aims to generate a list of deserving scholarship candidates, accompanied by a detailed analysis of the accuracy achieved by each technique employed in this study.*

I. Introduction

This paper explores various Machine Learning and Data Mining techniques applied to predict scholarship outcomes. It discusses different methods within these fields, detailing both their applications and encountered challenges. The merits and drawbacks of these approaches are thoroughly examined, culminating in the identification of the most effective method for predicting scholarship recipients.

Scholarships play a crucial role in providing financial assistance to students pursuing further education, typically offered by governmental or non-governmental organizations. Recognition through scholarships boosts students' confidence to pursue their future aspirations.

Common scholarship types include merit-based awards, those specific to certain students, and those tailored to particular career paths. These scholarships are awarded based on varying eligibility criteria.

For instance, academic scholarships often use Grade Point Average (GPA) as a key criterion for selection, while athletic scholarships consider performance across subjects, club involvement, community service, and more. This paper specifically addresses the application of Machine Learning and Data Mining in predicting scholarship recipients based on parameters such as percentage, GPA, marks, annual income, and communication skills.

II. Literature Survey

Thalia Anagnos (2018) emphasizes the critical importance of scholarships for supporting students financially and motivating them personally despite potential challenges like health or financial issues that could jeopardize their opportunities. Anagnos suggests that effective communication between teachers and industry professionals enhances program outcomes. The proposed system aims to provide essential support and programs to empower upper-class

engineering students to complete their degrees with the requisite skills, knowledge, and leadership qualities. While highlighting the significant impact of scholarships on recipients' leadership skills, attitudes, and university experiences, the study notes that the system's data processing time is prolonged, affecting efficiency.

Angela R. Bielefeldt (2015) examines the distinctions between engineering faculties' culture and the scholarship of teaching and learning (SOTL) within the engineering sector. Bielefeldt's research underscores the unique characteristics of faculty engaged in these activities across various categories such as assistant professors, full professors, and gender representation, particularly within Baccalaureate and Master's institutions. However, the study points out that the proposed system necessitates more extensive data generation and employs fewer predictive parameters.

In their 2017 paper, Sercan, Saaatci, Hande Cansiz, Gulsha Aslan, and Erkan Ozhan detail an artificial intelligence-based scholarship DGG credit pre-assessment system. This system distributes scholarships to eligible students across diverse institutions based on multiple criteria, although identifying deserving students remains challenging and time-consuming due to the reliance on question-and-answer-based prediction methods.

Okfalisa, Ratik a Fitreani, and Yelfi Vitreana (2018) explore data mining as a tool for predicting scholarship recipients, employing techniques such as K-Nearest Neighbors (KNN) and linear regression. They analyze key parameters like semester attendance, GPA, and various documentation to train and test their models, finding KNN to be more effective and efficient compared to linear regression.

Jonalyn Joy B. Labayne, Luster L. Mercado, and Jheanel Espiritu Estrada (2018) discuss FAITH institution's scholarship criteria, requiring a minimum GPA of 90% across all subjects. They compare decision tree, Naive Bayes, and K-NN algorithms to predict which students will maintain eligibility for their scholarships, using RapidMiner for data preprocessing. Decision tree emerges as the most accurate algorithm, highlighting subjects like algebra and English as critical factors influencing scholarship retention.

Moreover, current trends in industries such as e-commerce, tourism, and entertainment underscore the effective use of data mining, machine learning, and sentiment analysis to enhance decision-making processes and predict outcomes.

These methodologies play pivotal roles in analyzing consumer behavior, optimizing business strategies, and forecasting trends across various sectors.

III.System Analysis

Existing System: Scholarships provide crucial financial assistance for students to pursue higher education, often offered by governmental or non-governmental organizations to recognize student achievements.

This recognition boosts students' confidence in pursuing their goals. Scholarships are categorized into merit-based, student-specific, and career-specific types, tailored to different eligibility criteria. The current system employs the Random Forest algorithm for prediction, but it faces challenges like generating required data and using a limited number of parameters for accurate predictions. Disadvantages of Existing System:

- Relies solely on machine learning algorithms.
- Limited to regression techniques, which can reduce accuracy and increase complexity.

- Uses a smaller set of parameters for prediction.

Proposed System:

The proposed system integrates both machine learning (ML) and deep learning (DL) algorithms, leveraging ML's role in Artificial Intelligence for training data analysis across various fields like image processing and medical applications. ML algorithms predict unknown factors using training data inputs and outputs. Algorithms such as k-nearest neighbor, Decision tree, Support Vector Machine (SVM), Artificial Neural Network, and Naïve Bayes are employed to enhance scholarship prediction accuracy. The system aims to support upper-class students in completing engineering degrees with essential skills, knowledge, and leadership qualities, thereby maximizing scholarship impact and fostering student development.

Advantages of Proposed System:

- Implements both ML and DL algorithms for improved accuracy.
- Naïve Bayes algorithm offers good accuracy and flexibility.
- Streamlines scholarship prediction through ML and DL techniques, ensuring effectiveness.

This approach not only enhances prediction accuracy but also supports broader student development goals through enhanced algorithmic capabilities and data handling techniques.

This system analysis outlines the framework for implementing advanced data mining and machine learning techniques to predict student scholarship recipients effectively. By leveraging these technologies, educational institutions can optimize resource allocation and support students more efficiently in achieving their academic goals.

IV.System Study

Feasibility Study: In the system study phase for "A Review on Data Mining and Machine Learning Methods for Student Scholarship Prediction," a feasibility analysis is crucial to ensure the proposed project is viable and beneficial to stakeholders. The study encompasses three key aspects:

Economical Feasibility: Economical feasibility examines the financial impact of developing and implementing the system for predicting student scholarships. It evaluates whether the benefits derived from the system justify the costs incurred. Given that most technologies required for data mining and machine learning are freely available, the project remains within budgetary constraints. Customized solutions may necessitate minimal expenditures, which are justified by the potential long-term benefits of optimizing scholarship allocation and enhancing educational outcomes.

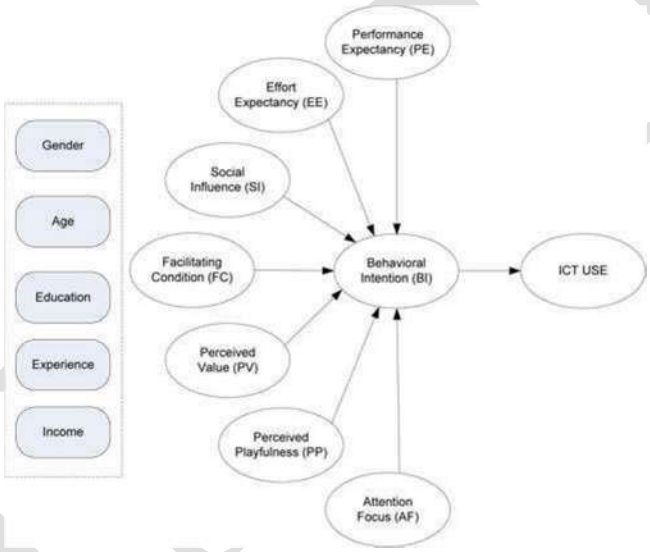
Technical Feasibility: Technical feasibility assesses whether the proposed system can be developed and integrated without imposing excessive demands on existing technical resources. The system must operate efficiently with minimal infrastructure changes. Utilizing widely adopted programming languages like Python and accessible frameworks such as scikit-learn ensures compatibility and ease of integration. This approach minimizes technical overhead and facilitates seamless deployment across educational institutions of varying scales.

Social Feasibility: Social feasibility evaluates the acceptance and usability of the system by stakeholders, primarily educational administrators, faculty, and students. User acceptance is critical for the system's success and involves comprehensive user training to ensure efficient utilization. Training programs will familiarize users with system functionalities and foster confidence in its capabilities. Encouraging user feedback and incorporating suggestions during system development ensures that the final product meets user expectations and enhances their confidence in utilizing predictive analytics for scholarship management.

By addressing economical, technical, and social feasibility aspects comprehensively, the system study phase ensures that the proposed project for predicting student scholarships through data mining and machine learning methods is not only feasible but also beneficial for enhancing educational support and outcomes.

V. System Design

System Architecture:

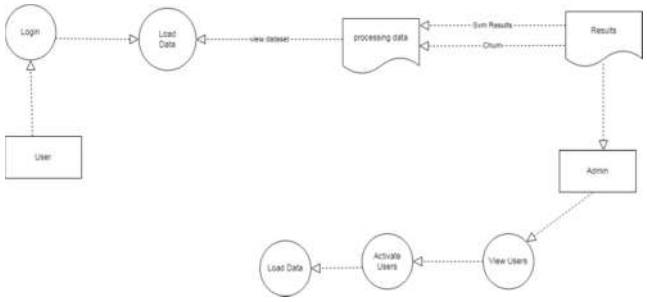


Data Flow Diagram: The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

The data flow diagram (DFD) is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



UML Diagrams

UML stands for Unified Modelling Language. UML is a standardized general-purpose modelling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artefacts of software system, as well as for business modelling and other non-software systems.
The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

Goals:

The Primary goals in the design of the UML are as follows:

Provide users a ready-to-use, expressive visual modelling Language so that they can develop and exchange meaningful models.

Provide extendibility and specialization mechanisms to extend the core concepts.

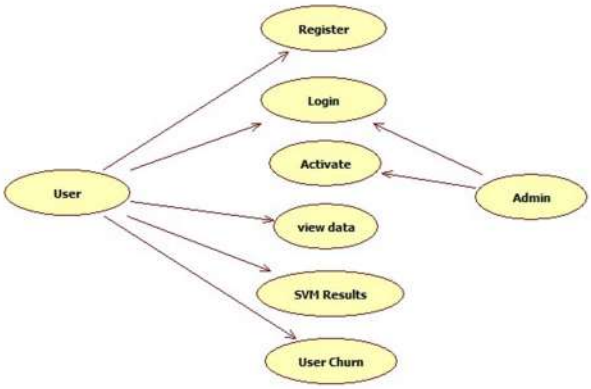
Be independent of particular programming languages and development process.

Provide a formal basis for understanding the modelling language.

Encourage the growth of OO tools market.

Support higher level development concepts such as collaborations, frameworks, patterns and components.

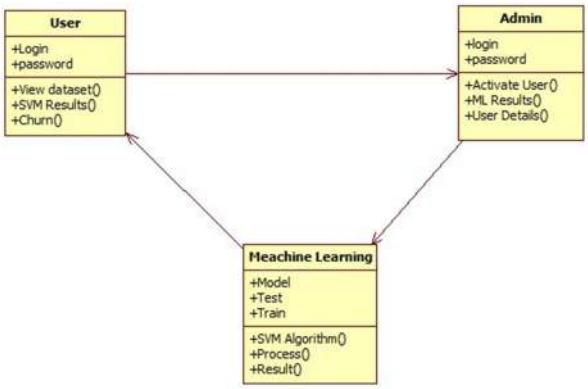
Integrate best practices.



A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

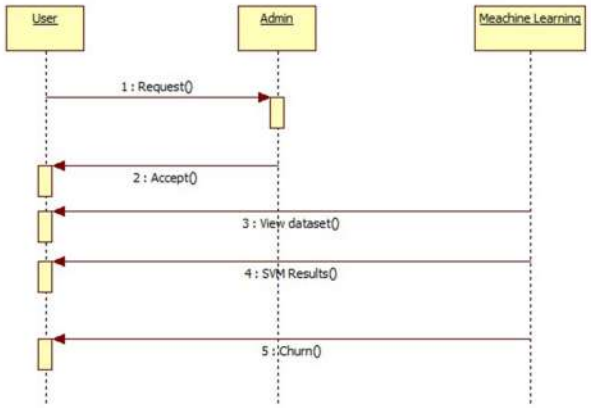
Class diagram:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



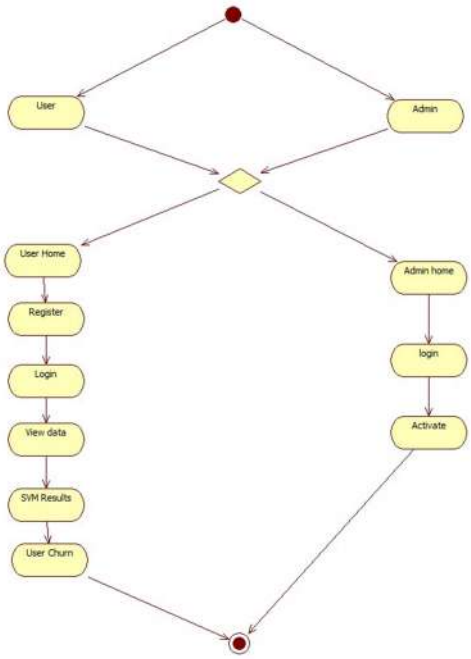
Sequence Diagram:

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams



Activity Diagram:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



VI.Modules Description

User:

The User module facilitates registration and activation for students participating in the scholarship prediction system. Users register with valid email and mobile details and await activation by the admin. Once activated, users can log in to upload datasets specific to scholarship prediction criteria. Data integrity is crucial, with requirements

that data be formatted correctly for machine learning algorithms, typically requiring numerical inputs or categorical data appropriately encoded. Users can also interact with the system to add new data entries or update existing ones, ensuring the predictive models are continually refined.

Admin:

Administrators possess elevated privileges, managing user activations and overseeing system operations. Admins facilitate user registrations, activate accounts, and monitor overall system functionality. They have access to comprehensive system data through a browser interface, enabling oversight of algorithmic results such as Accuracy, F1-Score, Precision, and Recall. Admins play a critical role in maintaining data integrity and ensuring the accuracy of predictive models used for scholarship allocation.

Data Preprocessing:

Data Preprocessing serves as a foundational module, preparing datasets for effective analysis and prediction. Techniques include cleaning data to remove noise, handling missing values through imputation or exclusion strategies, and transforming data into formats suitable for machine learning algorithms. Attribute selection and feature engineering may also be part of this module, optimizing datasets to enhance the accuracy and reliability of scholarship prediction models.

Machine Learning:

The Machine Learning module employs state-of-the-art algorithms to predict scholarship recipients based on prepared datasets. Techniques such as Decision Trees, Support Vector Machines (SVM), and Ensemble Methods are utilized to analyze historical data and extract patterns indicative of scholarship eligibility. The module facilitates model training on a portion of the data (e.g., 70%) and validation on the remaining portion (e.g., 30%), assessing model performance using metrics like Accuracy, Precision, Recall, and F1-Score. The goal is to deploy robust predictive models that assist educational institutions in effectively allocating scholarships to deserving students.

This structured approach ensures that the system for predicting student scholarships integrates seamlessly with data mining and machine learning methodologies, fostering accurate predictions and informed decision-making in educational settings..

VII. Conclusion

This paper presents a comprehensive survey of Machine Learning (ML) and Data Mining (DM) techniques employed in the scholarship prediction process, highlighting their effectiveness and ease of use. Predicting scholarship outcomes hinges on analyzing numerous parameters, necessitating robust training datasets crucial for ML and DM methodologies. However, the challenge lies in acquiring high-quality datasets and determining the optimal algorithm for accurate prediction. Despite these challenges, ML and DM methods enable systematic tracking of scholarship distributions and provide insights into the challenges faced by recipients in their academic

pursuits. Through these technologies, institutions can enhance their scholarship allocation strategies, ensuring more targeted and supportive educational assistance.

VIII. References

[1] Thalia Anagnos, Eva Schiorring, "Enhancing the Benefits of an Engineering Scholarship Program," IEEE, 2018.

[2] Angela R. Bielefeldt, "Attributes of Engineering Faculty Engaged in Teaching and Learning Scholarship," IEEE, 2015.

[3] K. Bunkar, "Data Mining: Predicting Performance Improvement of Graduate Students," IEEE, 2012.

[4] Sercan, Saatci, Hande Cansiz, Gulsha Aslan, Erkan Ozhan, "Artificial Intelligence-based Scholarship and Credit Pre-assessment System," IEEE, 2017.

[5] G. Upton, "Impact of Merit-Based Scholarships on Educational Outcomes," Journal of Labor Research, vol. 37, no. 2, pp. 235-261.

[6] Okfalisa, Ratik a Fitriani, Yelli Vitreana, "Comparison of Linear Regression and k-Nearest Neighbors for Scholarship Prediction," IEEE, 2018.

[7] N. T. Nghe, "Comparative Analysis of Techniques for Predicting Academic Performance," IEEE, 2007.

[8] Jonalyn Joy B. Labayne, Lester L. Mercado, Jheanel Espiritu Estrada, "Model Development of Student Scholarship Status at First Asia Institute of Technology and Humanities (FAITH)," IEEE, 2018.

[9] Suma, V., and Shavige Malleshwara Hills. "Data Mining for Demand Prediction in the Indian Market for Refurbished Electronics," Journal of Soft Computing Paradigm (JSCP), vol. 2, no. 2, 2020, pp. 101-110.

[10] Kumar, T. Senthil. "Data Mining Based Marketing Decision Support System Using Hybrid Machine Learning Algorithm," Journal of Artificial Intelligence, vol. 2, no. 3, 2020, pp. 185-193.

[11] Chakrabarty, Navoneel. "Regression Approach to Distribution and Trend Analysis of Quarterly Foreign Tourist Arrivals in India," Journal of Soft Computing Paradigm (JSCP), vol. 2, no. 1, 2020, pp. 57-82.

[12] Mitra, Ayushi. "Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)," Journal of Ubiquitous Computing and Communication Technologies (UCCT), vol. 2, no. 3, 2020, pp. 145-152.

IJMRR

[13] Hur, Minhoe, Pilsung Kang, and Sungzoon Cho. "Box-office Forecasting Based on Sentiments of Movie Reviews and Independent Subspace Method," Information Sciences, vol. 372, 2016, pp. 608-624.

[14] Vohra, S. M., and J. B. Teraiya. "Comparative Study of Sentiment Analysis Techniques," Journal JIKRCE, vol. 2, no. 2, 2013, pp. 313-317.

[15] Parvathy G, Bindhu JS. "Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media: A Review," Int J Comput Appl, vol. 134, no. 14, 2016, pp. 1-4.

IJMRR