Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Ensemble Machine Learning Approach for Identifying Determinants of Student Satisfaction

Priyadarshini P. [1] & K. T. Veeramanju [2]
[1] Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore – 575001, Karnataka India, ORCIDID: 0000-0003-4658-3742; Email: priyadarshini.pnair@gmail.com
[2] Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore – 575001, Karnataka India, ORCIDID: 0000-0002-7869-3914; Email: veeramanju.icis@srinivasuniversity.edu.in

ARTICLE INFO

Received: 18 Sep 2024
Revised: 10 Nov 2024
Accepted: 28 Dec 2024

ABSTRACT

This research focus on a comprehensive data-driven approach to analyzing student satisfaction using both unsupervised and supervised machine learning techniques. A dataset of 1,000 university students, including 38 survey-based features across academic experience, infrastructure, and career services, was utilized. Principal Component Analysis (PCA) was employed for dimensionality reduction. Clustering techniques including K-Means, DBSCAN, and Hierarchical Clustering identified distinct satisfaction profiles. K-Means (k=3) delivered the most interpretable structure and was selected for subsequent cluster profiling. A Random Forest classifier trained on normalized features achieved a 96% prediction accuracy, with F1-scores ranging from 0.94 to 0.97. The study culminates in targeted recommendations for institutional strategy based on cluster characteristics, illustrating the utility of ensemble learning in educational analytics.

Keywords: Clustering Analysis, K-Means, DBSCAN, Hierarchical Clustering, Random Forest classifier

I. INTRODUCTION

Student satisfaction is widely recognized as a key performance indicator (KPI) in the evaluation of higher education institutions. It serves not only as a proxy for the quality of academic and support services but also as a strong predictor of student retention, graduation rates, and alumni advocacy. In the current landscape of outcome-based education and accreditation-driven accountability, institutions are increasingly required to demonstrate continuous improvement through measurable outcomes.

Traditional methods of assessing student satisfaction such as anecdotal feedback or manual survey reviews are limited by subjectivity, lack of scalability, and delayed responsiveness. In contrast, data-driven approaches offer a more systematic and objective methodology for interpreting student feedback. These approaches leverage structured datasets, such as large-scale surveys, to uncover latent patterns that may not be visible through conventional analysis. Student satisfaction is one of the crucial elements of showing the quality of educational institutions. Traditionally the student satisfaction was measured using a single question or some questions which gave only yes or no response, which may not capture the complexity of their overall experience (Salameh, M.,Touqan, B., &Suliman, A. (2024). Higher Education Institutions specifically engineering and its allied branches are functioning in a competitive environment; they require continuous improvements to progress student satisfaction and maintain academic excellence. Understanding students' satisfaction and identifying areas of dissatisfaction can improve the educational experiences (Clemons, R., & Jance, M. 2024)). In today's competitive environment, every organization must concentrate on their customer satisfaction and feedback. From a systems thinking perspective, academic institutions also function similarly to any other organizations, in which their long-term success is strongly correlated to student satisfaction. As a result, enhancing student satisfaction has become a crucial part of every educational institution (Kuzehgar, M., & Sorourkhah, A. 2024). Student satisfaction is influenced by both academic and administrative

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

services, with non- academic factors playing a major part in shaping their overall satisfaction towards the institution (Ruranga, 2024).

Educational institutions can effectively use clustering results to better understand student needs, prioritize service improvements, and design more targeted interventions (Maulidya, A., et al 2024). Clustering techniques, particularly K-Means and related unsupervised learning algorithms, are increasingly employed in educational data mining to uncover latent patterns in student behavior and satisfaction. By grouping students based on variables such as academic performance, engagement levels, or feedback responses, clustering helps institutions identify distinct profiles such as high achievers, average performers, and at-risk students. For example, clustering based on metrics like study hours, attendance, and tutoring sessions can reveal engagement trends that support early interventions. Similarly, satisfaction-based clustering can guide personalized support strategies and infrastructure planning. These methods not only enhance the interpretability of complex survey data but also allow educators and policymakers to prioritize resources effectively and respond proactively to students' diverse needs (Durachman, Y., & Rahman, A. W. B. A. 2025). Clustering techniques are widely used in educational data mining to group students based on academic behavior and engagement patterns. This helps identify student profiles such as high achievers, average performers, and those needing support, enabling more targeted interventions. Studies have shown that clustering, when combined with predictive modeling, can significantly enhance decision-making in personalized education and academic performance improvement (Claudio, B. M. (2024). Clustering analysis has proven valuable in identifying academic performance patterns among students, enabling targeted educational support (Maguate, G. (2024). In addition to clustering, predictive models such as Random Forest, Support Vector Machines (SVM), and Neural Networks are commonly used to forecast student performance. These models can uncover patterns in academic records, engagement, and attendance to help educators intervene early. (Madahana, M. C., et al, 2024) demonstrated that machine learning techniques can effectively support student retention by identifying those at risk of dropping out. SVM and decision tree-based models perform well in classification tasks, highlighting their usefulness in educational prediction contexts.

In this study, survey data collected from 1,000 university students comprising 38 variables related to academic quality, campus infrastructure, extracurricular involvement, and career services was analyzed using advanced machine learning techniques. The methodology employed an ensemble framework of both unsupervised and supervised learning algorithms. By using clustering algorithms (e.g., K-Means, DBSCAN), the study segmented students into distinct satisfaction profiles, and subsequently trained a predictive model (Random Forest classifier) to generalize these patterns across new data. This dual approach supports evidence-based policy formulation and enhances the institution's ability to deliver targeted academic interventions and personalized student support. This paper structured into five SECTIONS including introduction. The next section gives the detailed literature survey of existing study. The third section explains the methodology of the study in detail followed by the result and analysis of the study. And the last section gives the conclusion.

II. RELATED WORK:

(Aulakh, K., et al, 2023) emphasized the foundational role of Educational Data Mining (EDM) in transforming raw student data into actionable insights. EDM integrates diverse methodologies to uncover meaningful academic patterns that help in improving student learning outcomes and institutional performance. These insights are then refined using various machine learning techniques to develop responsive, data-driven learning environments. According to (Helm, J. M., et al, 2020) rapid technological advancements and machine learning has become increasingly relevant in educational settings. The increased availability of large datasets and improved computational power have shifted focus from basic pattern recognition to sophisticated models like deep learning, enabling more accurate predictions and real-time analysis in education. (Suryadevara, C. K. 2018) explored a grammar-guided genetic programming algorithm (G3PMI) to predict academic success with an accuracy of 74.29%. Other studies developed predictive models using student attendance and previous subject scores, showing improved performance with larger datasets and reaching over 70% accuracy using neural networks. (Talwar, S.,et al, 2021) applied Artificial Neural Networks (ANNs) to predict student exam performance, achieving an impressive 85% accuracy. (Kotsiantis et al., 2010) compared several machine learning techniques, concluding that the Naïve Bayes algorithm achieved an average accuracy of 73%, making it a reliable yet accessible tool for academic forecasting. (Hasan, H. R., et al, 2019)

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

reviewed 29 studies on student performance prediction highlighted six major ML models they are decision trees, ANN, SVM, KNN, linear regression, and Naive Bayes. Among these, ANN models consistently outperformed others in terms of accuracy. The findings confirm the growing research interest in this area and the variety of ML techniques applied. (Bhutto, E. S., et al, 2020) in his study found that supervised machine learning algorithm Sequential Minimal Optimization, performed better than logistic regression in predicting student academic performance. The results shows that SMO achieves higher accuracy and helps identify key influencing factors such as teacher performance and student motivation, which can support early interventions to reduce student dropout rates. (Albreiki, B., et al, 2021) found that Educational Data Mining (EDM) helps improve the learning environment by using machine learning and data analysis tools. Their review showed that many studies use student data to predict which students are at risk. Delen, D. (2010) study showed in which students might drop out can help colleges keep more students. The researcher looked at five years of student data and found that using an ensemble worked better than using single model to make predictions. (Agrawal, H., & Mavani, H. 2015) proposed a model to predict the performance of students in an academic organization. The algorithm employed is a machine learning technique called Neural Networks. The study of Almarabeh, H. (2017) tested different data mining approaches to see which model predicts student performance. Using Naive Bayes, Bayesian Network, ID3, J48, and Neural Network the researchers identified that the Bayesian Network gave the accurate results. (Sekeroglu, B., et al, 2019) showed that student performance can be predicted and classified using machine learning. This shows that using technology like machine learning can help improve how we understand and support student learning.

Based on the literature review conducted many researchers have used data mining and machine learning approaches to analyze the student satisfaction. These studies and approaches helped to identify students who need support and attention through that learning strategies can be improved. Most of the studies focus on student performance in academics. Less studies are conducted to identify the important features that affect student satisfaction by analyzing survey responses. Limited features are considered for the satisfaction analysis. All aspects of satisfaction levels should be considered. Most of the studies focus only on one method either grouping students based on satisfaction or just predicting the outcome. To improve the results an ensemble approach is identified and tested in this study. Initially grouped students based on their satisfaction levels using algorithms like K-Means, and then trained a model to predict those groupings using Random Forest. This approach was chosen because it combines the strengths of both techniques and helps make decisions based on real student feedback. It gives a more complete view of student satisfaction and can help universities to take a better decision for the improvement. The next section elaborates the data collection methods for the proposed study.

III. METHODOLOGY

Dataset Overview:

An online questionnaire format is designed for the survey to get valuable insights from students on their satisfaction towards the institution. To prepare the survey questions, various attributes which is contributing students' satisfaction are included. The questions are prepared after the discussion with various students and academicians for the better results and decision making. English language is used with a user-friendly online platform, which ensures accessibility to all the students selected for the survey. The survey questions address multiple independent variables and one dependent variable. The Likert scale method is used to measure the responses. Which provides a structured way to collect data from participants at their level of agreement or disagreement with a series of questions. Values for the scale are represented in numerical values, like a 5-point scale. The questionnaire consists of 38 questions reflecting different aspects of student satisfaction such as demographic information, academic experience, faculty and mentorship, facilities and infrastructure, placement counseling, extracurricular activities, overall satisfaction, and one open-ended question. The dependent variable is willing to recommend the institution to others. Academic experience, Faculty and mentorship, facilities and infrastructure, placement counseling, extracurricular activities are the various independent variables. The population in this study 1000 students from a private university. Permission to distribute an online questionnaire was obtained from the respective institution and the methods also informed clearly. The purpose of the study was clearly communicated to the participants. data collected from graduation as well as post-graduation students. The survey was conducted through online Google form for the better reachability.

Data Preprocessing:

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In the data preparation phase, the initial task was the thorough cleaning and standardization of column names to rectify typographical errors, formatting inconsistencies, and ambiguous feature labels. This step ensured that all feature names were syntactically uniform, semantically meaningful, and compatible with downstream machine learning operations. Proper feature labeling is critical in data-driven studies, not only to facilitate interpretation but also to avoid potential runtime errors during automated analysis.

Subsequently, the dataset underwent normalization using the Min-Max scaling technique. This method transformed each feature to a common numerical range [0, 1], calculated using the equation (1)

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \text{----------} (1)$$

where X is the original feature value, and X_{min} and X_{max} denote the minimum and maximum values of that feature, respectively. Normalization was a necessary pre-processing step to ensure that no single feature with a larger magnitude disproportionately influenced distance-based models such as K-Means clustering. It also promoted stable convergence during model training.

Following normalization, dimensionality reduction was performed using PCA, a linear transformation technique that projects high-dimensional data into a lower-dimensional subspace while preserving the maximum possible variance. PCA identifies orthogonal axes, known as principal components, along which the variance in the data is maximized. This transformation is particularly advantageous for high-dimensional datasets with potentially correlated or redundant features, as it condenses the dataset into a few composite variables without significant loss of information.

In this study, the original 38-dimensional feature space was reduced to two principal components. The first principal component (PCA1) captured approximately 18.94% of the total variance, representing the most significant linear combination of features. The second component (PCA2), orthogonal to the first, explained an additional 7.28% of the variance. Together, these two components accounted for 26.22% of the total variation within the dataset. Although this may seem modest, in the context of educational survey data where many variables are interrelated this level of variance retention is sufficient for visualization and preliminary clustering.

The two-dimensional PCA projection not only facilitated intuitive graphical visualization of the student clusters but also improved the effectiveness of clustering algorithms by eliminating redundant and noisy feature dimensions. This step was instrumental in uncovering latent satisfaction profiles within the student population.

Clustering Techniques:

To identify various satisfaction groups from the student data clustering algorithms are identifies. They are K-Means, DBSCAN, and Hierarchical Clustering. Each algorithm was applied to the PCA-transformed dataset, which reduced the original 38-dimensional survey features to two principal components.

K-Means clustering partitions the dataset into k groups by minimizing the distance between data points and their respective cluster centroids. It was selected for its simplicity, efficiency, and high interpretability. The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was also evaluated to detect clusters of varying density and to identify potential outliers. In addition, Hierarchical Clustering was used to visualize nested clusters and understand relationships among student responses at multiple levels of granularity.

To determine the most suitable algorithm, Silhouette Score was used as the internal validation metric. This score measures how similar each point is to its own cluster compared to other clusters. Among the tested methods, K-Means with k=3 achieved the highest silhouette score of approximately 0.09, which, although modest, was acceptable given the noisy nature of survey data. The resulting three clusters corresponded logically to highly satisfied, moderately satisfied, and dissatisfied or at-risk students. Based on its performance and clear group separation, K-Means was selected as the final clustering method.

Predictive Modeling:

After the completion of clustering process, a Random Forest classifier was trained to predict student satisfaction groups using the original 38 survey features as input and the K-Means cluster labels as target classes. Random Forest

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

is an ensemble learning method that builds multiple decision trees and combines their predictions for higher accuracy and robustness against overfitting.

The dataset was split into training and testing sets using an 80:20 ratio. Standard hyperparameters such as the number of estimators (n_estimators = 100), maximum tree depth, and minimum samples per leaf were applied. Grid search and cross-validation were optionally used to fine-tune parameters, although default settings yielded high performance.

The model is evaluated using the metrics Accuracy, F1 score, Precision and Recall. The high accuracy and interpretability of Random Forest made it a suitable choice for predicting satisfaction profiles based on student feedback data. This study used both clustering and prediction techniques to better understand student satisfaction. The combination of these two methods provides a strong and reliable way to group students based on their feedback and to predict satisfaction levels using their survey responses. The following section presents the outcomes of the clustering and classification processes and highlighted how effectively the proposed methods grouped the students and predicted their satisfaction categories and extracted the most important features affecting satisfaction. With the proposed methodology, the next section presents the results obtained from clustering and predictive modeling, highlighting the effectiveness of the approach in identifying distinct student satisfaction profiles and predicting satisfaction levels with high accuracy.

IV. Results

The result section explains the outcomes of the clustering and prediction processes applied to the student satisfaction survey data. The aim is to group students with similar satisfaction patterns and then accurately predict these groupings using a classification model. The results are discussed in two parts. First, the findings from the clustering analysis to identify distinct satisfaction groups and second, the performance of the predictive model in classifying students based on their responses.

Unsupervised Clustering Analysis:

Three clustering algorithms K-Means, DBSCAN, Hierarchical Clustering were tested. The highest silhouette score appears around k=2 or k=3, though still relatively low overall. After k=4, the scores tend to decrease, indicating diminishing returns in cluster quality with more groups. Table 1 describe the use of each clustering algorithms used.

Table 1: algorithm used and description and its use

Algorithm	Description	Use
K-Means	Partitions data into k clusters by minimizing intra-cluster variance	To group students based on similar satisfaction patterns
DBSCAN	Density-based clustering, good for discovering outliers	To detect isolated/dense satisfaction groups
Hierarchical	Builds a tree of clusters based on distances	To visualize student similarity at different granularities

The core objective of applying clustering in this study was to uncover latent student satisfaction segments based on survey responses. Clustering is an unsupervised learning method that groups data points such that, Intra-cluster similarity is maximized (students within a group are similar). Inter-cluster dissimilarity is maximized (students from different groups are distinct).

Among various clustering methods tested (K-Means, DBSCAN, Hierarchical), K-Means clustering with a fixed number of clusters k emerged as the most interpretable and efficient method.

The Silhouette Score is a widely used internal validation metric that quantifies how well each data point fits within its assigned cluster compared to other clusters.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

For each point i

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

Where, $a(i)$ average intra-cluster distance (cohesion) and $b(i)$ average nearest-cluster distance (separation)

The score ranges from -1 to +1**.** Perfect clustering (well-separated and cohesive), ~0**:** Overlapping clusters and less that 0: Incorrect clustering (point is closer to another cluster). This Silhouette analysis provides an objective criterion to determine the optimal number of clusters without relying on visual inspection alone. K-Means with k=3 was selected based on silhouette analysis (score ≈ 0.09). The highest silhouette score (~0.09) occurred at k=3, indicating that this configuration best balanced compactness and separation**.**

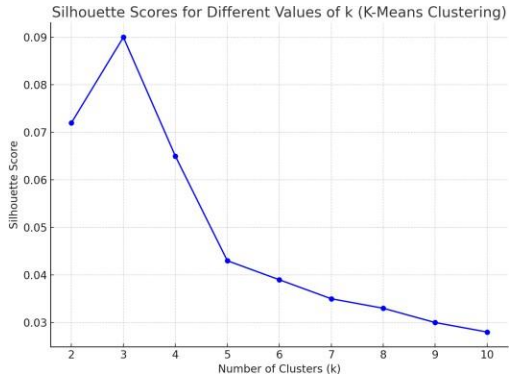

Figure 1. Silhouette scores for different cluster counts (k), indicating optimal value at k = 3

Although a silhouette score of 0.09 is relatively low in absolute terms, it is acceptable in real-world social science and survey data, where:Human behavior data is noisy and non-spherical**.**Satisfaction scores often overlap between groups.There is no ground truth for how many types of students should exist.In practice. Cluster 0: Highly satisfied students. Cluster 1**:** Moderately satisfied students and Cluster 2**:** Dissatisfied or at-risk students. This three-group segmentation aligned logically with expected satisfaction tiers in education settings, thereby reinforcing the selection of k = 3 both empirically and semantically**.**

K-Mean (with k=3) is the most interpretable, Mathematically supported by silhouette analysis, Aligned with domain-specific expectations. Hence, despite modest silhouette scores, K-Means with k=3 was selected.

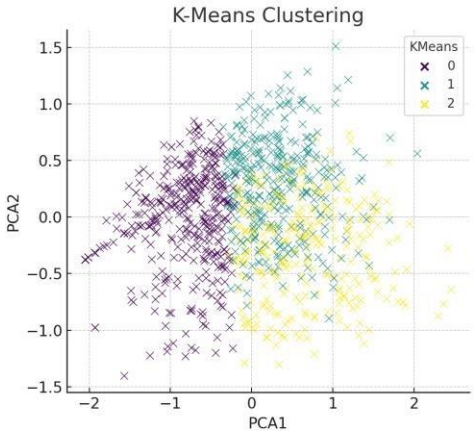

Figure 2: PCA Clustering Visualization of K-mean

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

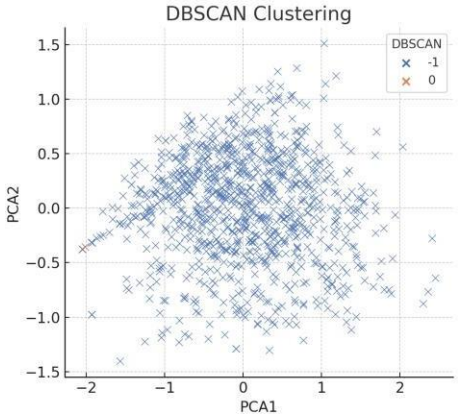

Figure 3: PCA Clustering Visualization of DBSCAN

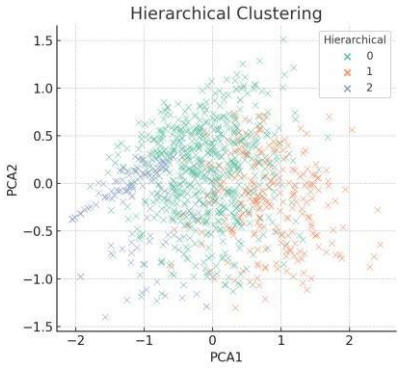

Figure 4: PCA Clustering Visualization of hierarchical Clustering

Cluster Profiling:

Each cluster was analyzed based on mean scores per feature: Cluster 0 (highly satisfied), Cluster 1 (moderately), and Cluster 2 (least satisfied). Radar charts illustrated in Figure 5 differences in satisfaction dimensions. The radar chart compares the average satisfaction levels across key areas for the three student groups formed using K-Means clustering. These areas include academic support, resource access, course relevance, event quality, infrastructure, employability services, and the overall recommendation of the institution. Students in Cluster 0 showed the highest satisfaction in nearly all areas. Cluster 1 had moderate satisfaction, while Cluster 2 reported the lowest levels, especially in course relevance and event quality. This visual comparison supports the idea that the clusters represent highly satisfied, moderately satisfied, and dissatisfied student groups. The chart helps in understanding the distinct needs of each group and provides useful insights for improving student services.

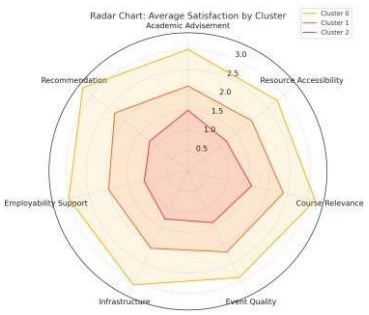

Figure 5. Radar chart comparing average satisfaction scores across clusters for key features.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

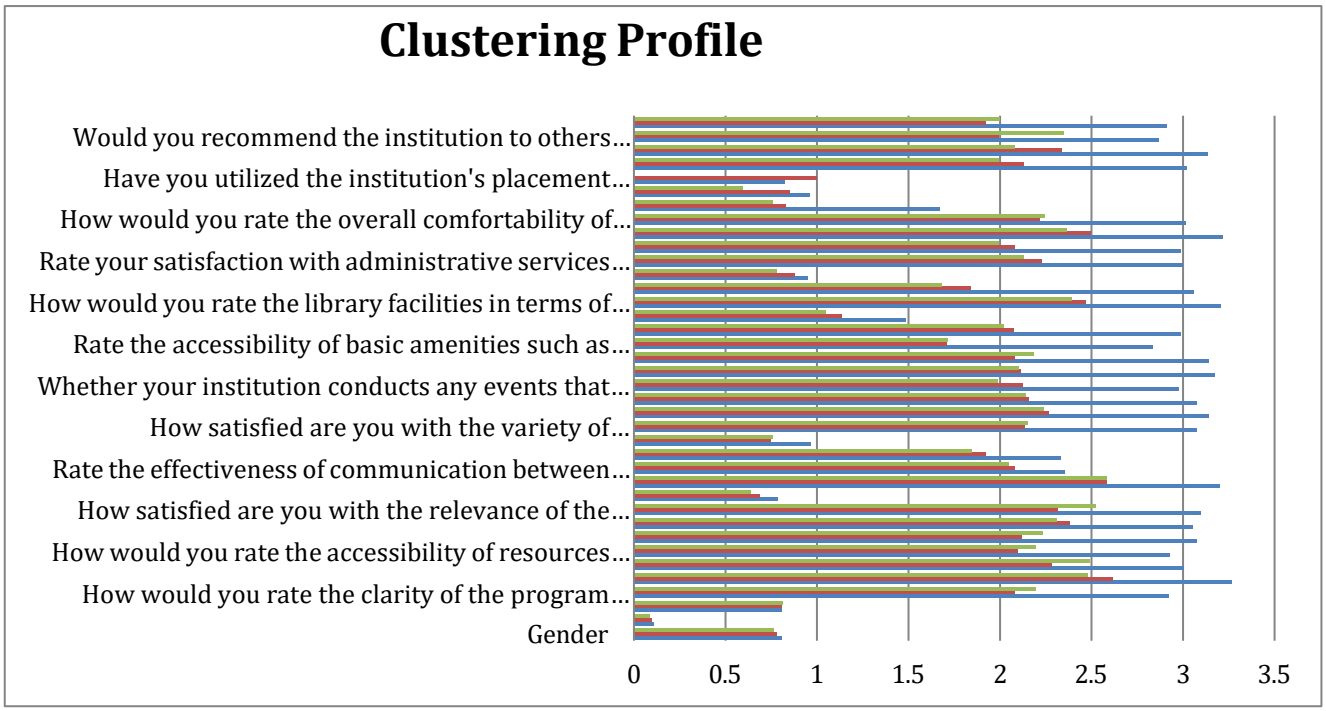

Figure 6. Average satisfaction scores for each survey feature across student clusters.

This dataset represents the average values of various survey-based features across three student satisfaction clusters, which were derived using K-Means clustering. This Cluster profiling is the process of summarizing and characterizing each group (cluster) identified by a clustering algorithm. In this case: Cluster 0**:** Represents one group of students (possibly the highly satisfied group) Cluster 1**:** Another distinct group (moderately satisfied). Cluster 2**:** A third group (least satisfied). Each row shows the mean score of a particular feature (e.g., academic support, gender ratio, course clarity) for the students within each cluster. Interpretation of Key Features

Demographics: Features like *Gender*, *Age*, and *Department* are normalized between 0 and 1. The small variation across clusters indicates similar demographic distributions.

Academic Experience: For example, the feature *"How would you rate the clarity of the program..."* shows a much higher score in Cluster 0 (2.92) than Cluster 1 (2.08) or Cluster 2 (2.19). This suggests that students in Cluster 0 are significantly more satisfied with program clarity.

Faculty Support and Advisement: Cluster 0 again leads with 3.27, compared to 2.61 (Cluster 1) and 2.48 (Cluster 2), reinforcing the notion that Cluster 0 represents the more satisfied students.

1. Insightful Segmentation: This profile helps institutions understand what distinguishes each group. For example: Cluster 2 might need intervention in academic advising. Cluster 1 could benefit from improvements in course relevance or infrastructure.
2. Data-Driven Decision-Making: These averages help prioritize institutional policies. You don't just know *that* students are unsatisfied — you know *where* and *how much*.
3. Personalized Strategies: For Cluster 2: Introduce mentorship and academic counseling. For Cluster 0: Retain satisfaction by involving them in feedback loops or peer-support roles.

Predictive Modeling

Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their outputs to improve predictive accuracy and reduce overfitting. It also offers feature importance rankings, making it valuable for identifying key satisfaction factors.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Following the clustering process, a supervised learning model was developed to predict a student's satisfaction group (Cluster 0, 1, or 2) based on the full set of 38 survey features. A Random Forest classifier was selected due to its robustness to overfitting, capacity to handle non-linear relationships, and ability to provide feature importance metrics, which are vital for model explainability.

The dataset was randomly split into training (80%) and testing (20%) subsets, and the model was trained using 10-fold cross-validation to ensure generalizability. Hyperparameter tuning was conducted using grid search to optimize the number of estimators, maximum depth, and minimum sample splits.

The final model achieved an overall accuracy of 96%, demonstrating excellent generalization capability. Performance metrics for each satisfaction cluster are summarized in Table 2.

Table 2: Precision, Recall, and F1-score for each predicted student satisfaction cluster.

Metric	Cluster 0	Cluster 1	Cluster 2
Precision	0.97	0.96	0.95
Recall	0.92	0.97	0.99
F1- score	0.94	0.97	0.97

A Random Forest model predicted satisfaction clusters using all features. Accuracy was 96%, and F1-scores were high across all clusters (0.94–0.97). Figure 7 illustrates the top 15 most influential features based on the Random Forest's feature importance scores. Key predictors included satisfaction with faculty advisement, clarity of program requirements, curriculum relevance to industry, and availability of career services. These features not only improved model accuracy but also provided institutional insight into factors driving satisfaction.

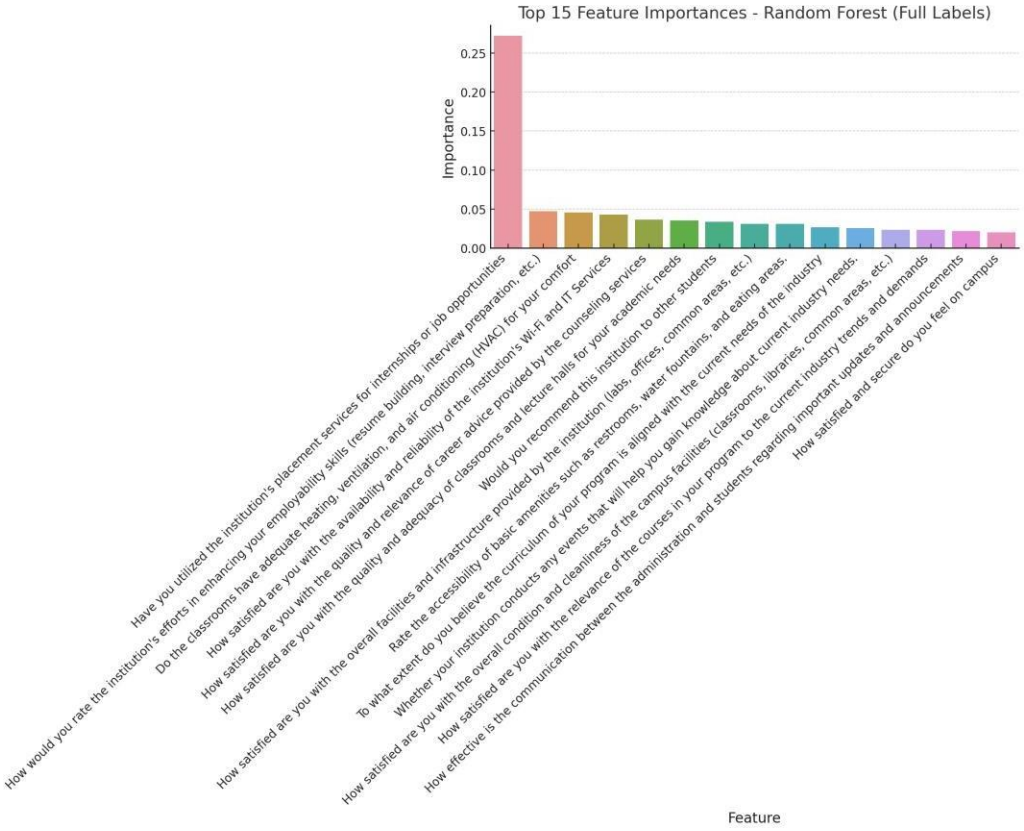

Figure 7. Top 15 most important features in predicting student satisfaction clusters.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

DISCUSSION

Each algorithm employed in this study was selected based on its compatibility with the nature of the dataset and the specific analytical objectives. Principal Component Analysis (PCA) was used as a dimensionality reduction technique to mitigate the curse of dimensionality and to enhance the performance of clustering algorithms. It projected the original 38-dimensional feature space into two principal components that retained sufficient variance for visual analysis and cluster separation. For clustering, K-Means proved to be the most effective approach for segmenting students into interpretable satisfaction groups. Despite the low absolute silhouette scores (common in social science datasets), K-Means yielded stable, semantically meaningful clusters. DBSCAN was tested to identify non-linear structures and potential outliers, which contributed to understanding density-based groupings, although it suffered from sensitivity to parameter tuning. Hierarchical clustering added hierarchical insights into how student profiles relate across different levels of granularity but lacked sharp separation in the dataset's latent structure. The use of the Random Forest classifier in the supervised learning phase was justified by its robustness, non-parametric nature, and ability to handle high-dimensional input without strong assumptions about the distribution. It also provided interpretable results through feature importance ranking and achieved excellent classification metrics (96% accuracy and F1-scores > 0.94 for all clusters). This interpretability was particularly valuable for linking student satisfaction categories back to actionable institutional factors. Based on the cluster profiling and classification outcomes, tailored recommendations were developed for each group to support evidence-based decision-making in higher education management:

Cluster 0 – Highly Satisfied Students: This group demonstrated high scores across most academic and infrastructural features. Institutions should maintain current levels of engagement and academic support for these students. In addition, they can be engaged as peer mentors, student ambassadors, or part of feedback committees, helping drive quality assurance through participatory leadership.

Cluster 1 – Moderately Satisfied Students: While reasonably content, this group expressed lower satisfaction in areas such as academic advising, course variety, and infrastructure quality. Targeted improvements in these areas — such as expanding specialization options or upgrading learning spaces — can convert moderate satisfaction into high satisfaction and prevent potential attrition.

Cluster 2 – Dissatisfied or At-Risk Students: This segment showed the lowest scores, particularly in areas like career support, faculty mentorship, and curriculum relevance. Immediate and sustained intervention is critical for this group. Suggested actions include implementing personalized academic counseling, career pathway mapping, and student success workshops. Retaining these students is not only a reputational concern but also financially prudent for the institution. These findings demonstrate that a combined clustering and classification approach not only enhances the understanding of student satisfaction patterns but also offers practical guidance for targeted institutional improvements.

IX. CONCLUSION

This study proposed and validated an interpretable, data-driven framework for analyzing and predicting student satisfaction using a combination of unsupervised clustering and supervised ensemble learning. The hybrid use of PCA for dimensionality reduction, K-Means for segmentation, and Random Forest for predictive modeling provided both analytical depth and practical utility.

By segmenting students into distinct satisfaction groups and predicting those with high accuracy, the study offers a powerful tool for institutional planning, targeted intervention, and policy formulation. Unlike traditional satisfaction studies limited to descriptive statistics, this machine learning-based approach enables proactive, scalable, and individualized support.

In essence, the integration of machine learning with educational data mining empowers universities to move beyond reactive measures, enabling strategic academic decision-making and continuous improvement in student experience.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

REFERENCES:

[1] Agrawal, H., & Mavani, H. (2015). Student performance prediction using machine learning. *International Journal of Engineering Research and Technology, 4(03), 111-113*.

[2] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences, 11(9), 552*.

[3] Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. *International Journal of Modern Education and Computer Science, 9(8), 9*.

[4] Aulakh, K., Roul, R. K., & Kaushal, M. (2023). E-learning enhancement through educational data mining with Covid-19 outbreak period in backdrop: A review. *International Journal of Educational Development, 101*, 102814.

[5] Bhutto, E. S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020, February). Predicting students' academic performance through supervised machine learning. In *2020 International Conference on Information Science and Communication Technology (ICISCT)* (pp. 1–6). IEEE.

[6] Claudio, B. M. (2024). Application of Data Mining for the Prediction of Academic Performance in University Engineering Students at the National Autonomous University of Mexico, 2022. *LatIA, (2), 4*.

[7] Clemons, R., & Jance, M. (2024). Defining Quality in Higher Education and Identifying Opportunities for Improvement. *Sage Open, 14(3)*.

[8] Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49(4), 498–506*.

[9] Durachman, Y., & Rahman, A. W. B. A. (2025). Clustering Student Behavioral Patterns: A Data Mining Approach Using K-Means for Analyzing Study Hours, Attendance, and Tutoring Sessions in Educational Achievement. *Artificial Intelligence in Learning, 1(1), 35–53*.

[10] Hasan, H. R., Rabby, A. S. A., Islam, M. T., & Hossain, S. A. (2019, July). Machine learning algorithm for student's performance prediction. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1–7). IEEE.

[11] Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., ... & Ramkumar, P. N. (2020). Machine learning and artificial intelligence: definitions, applications, and future directions. *Current Reviews in Musculoskeletal Medicine, 13, 69–76*.

[12] Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems, 23(6), 529–535*.

[13] Kuzehgar, M., & Sorourkhah, A. (2024). Factors affecting student satisfaction and dissatisfaction in a higher education institute. *Systemic Analytics, 2(1), 1–13*.

[14] Madahana, M. C., & Ekoru, J. E. (2024, October). Comparative Study of Machine Learning Algorithms for Student Retention, Early Warning and Intervention Systems for Institutions of Higher Learning. In *2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 366–371). IEEE.

[15] Maguate, G. (2024). Analyzing Student Performance: A Clustering Approach for Academic Intervention. *Available at SSRN 4873086*.

[16] Maulidya, A., Sitorus, Z., Siahaan, A. P. U., & Iqbal, M. (2024). Analysis Of Increasing Student Service Satisfaction Using K-Means Clustering Algorithm and Gaussian Mixture Models (GMM). *International Journal Of Computer Sciences and Mathematics Engineering, 3(1), 29–35*.

[17] Ruranga, C. (2024). Exploring higher education students' satisfaction for quality improvement: A case study of the African Centre of Excellence in Data Science. *International Journal of Education and Practice, 12(3), 719–729*.

[18] Salameh, M., Touqan, B., & Suliman, A. (2024). Enhancing student satisfaction and academic performance through school courtyard design: A quantitative analysis. *Architectural Engineering and Design Management, 20(4), 911–927*.

[19] Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019, March). Student performance prediction and classification using machine learning algorithms. In *Proceedings of the 2019 8th International Conference on Educational and Information Technology* (pp. 7–11).

[20] Suryadevara, C. K. (2018). Predictive modeling for student performance: Harnessing machine learning to forecast academic marks. *International Journal of Research in Engineering and Applied Sciences (IJREAS), 8(12)*.

[21] Talwar, S., Talwar, M., Tarjanne, V., & Dhir, A. (2021). Why retail investors traded equity during the pandemic? An application of artificial neural networks to examine behavioral biases. *Psychology & Marketing, 38(11), 2142–2163*.