ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2025.0522

STUDENT SATISFACTION ANALYSIS WITH GENETIC ALGORITHM-BASED DATA AUGMENTATION AND REGRESSION MODELS

P. Priyadarshini and K.T. Veeramanju
Institute of Computer Science and Information Science, Srinivas University, India

Abstract

Student satisfaction plays an important role in determining the quality, retention, and reputation of an institution. However, limited survey data can reduce the accuracy of predictive models. This study explores how Genetic Algorithm based data augmentation can improve dataset reliability and enhance analysis using LASSO and Ordinal Regression. By generating synthetic responses, GA expands the dataset while maintaining statistical accuracy, leading to better feature selection and ranking. LASSO Regression identified key factors influencing student satisfaction, such as career services, curriculum relevance, faculty support, and extracurricular activities, while Ordinal Regression pointed out that administrative inefficiencies negatively affect satisfaction levels. The results highlight that academic and career-related aspects have a greater impact on student satisfaction than infrastructure facilities. To enhance student experiences, institutions should focus on faculty mentorship, career counseling, and aligning the curriculum with industry needs. This study shows that GA-based data augmentation can significantly improve predictive modeling for student satisfaction analysis and offers practical recommendations for institutional development. Future research can incorporate machine learning techniques for more accurate predictions and tailored strategies to improve student success.

Keywords:

Student Satisfaction, Genetic Algorithm, Ordinal Regression, LASSO Regression

1. INTRODUCTION

Student satisfaction is an essential factor in evaluating the quality of educational institutions [1]. It reflects students' experiences with academic programs, faculty support, infrastructure, extracurricular activities, and administrative services. By identifying the factors affecting student satisfaction, institutions can concentrate areas for improvement and enhance the overall learning environment [2] . Manually identifying the factors which contribute to students' satisfaction is a difficult task. Regression analysis helps address this challenge by providing a quantitative approach to understanding the relationship between student satisfaction and various influencing factors [3]. By applying regression models, institutions can determine which factors such as faculty support, career counseling, curriculum, infrastructure or extracurricular activities have the most significant impact on student satisfaction. By extracting the features the institutions can make decisions which will help administrators to prioritize areas that require improvement.

Data collection in educational research is often cumbersome. To perform machine learning analysis a large amount of data is required. If the number of data is more then the model efficiency will improve [4]. Gathering large-scale and high-quality survey responses requires significant time and effort. At the same time limited data availability can reduce the accuracy of regression models. Small datasets often lead to biased predictions, reduced generalizability, and poor model performance. Data augmentation techniques are widely used to reduce this issue, which artificially increases the dataset size while preserving its originality [5].

There exist various data augmentation techniques in which genetic algorithms provide an efficient method for generating synthetic data [6]. The genetic algorithm is more suitable for survey data because of its functionality to maintain the originality of real data and maintain the existing pattern instead of adding random values [7]. While compared with other Artificial Intelligence models, Genetic Algorithm is simple [8]. Genetic Algorithms are commonly used to create effective solutions by using the methods inspired by biology, they are mutation, crossover, and selection [9].The newly generated data follows the logic of the original dataset. It is ensuring balanced distributions among different satisfaction levels.By using the augmented data using a genetic algorithm, the regression model will improve the accuracy for better feature extraction. The result of the regression identifies the factors influencing student satisfaction. Through this, institutions can make better decisions.

This paper is structured into four chapters. The Paper commences with the articulation of its objectives and the underlying motivation. It then progresses to evaluate the data, encompassing an examination of primary data sources. The subsequent section outlines the methodologies employed and the construction of predictive models. The penultimate chapter presents the findings and engages in a discussion about these outcomes. Finally, the paper wraps up with a critical review of the results and the effectiveness of the methodologies applied.

2. RELATED WORKS

Numerous studies have been conducted to analyze student satisfaction using statistical and machine learning techniques, emphasizing the importance of understanding the variables that significantly influence students' educational experiences. Traditional models, such as linear regression and decision trees, have been applied in educational research to predict student satisfaction and performance outcomes based on demographic, academic, and institutional variables [10]. However, these models often suffer from overfitting when applied to small datasets, which is a common limitation in education-related studies.

To address the issue of limited data availability, various data augmentation strategies have been proposed. Conventional methods such as SMOTE (Synthetic Minority Oversampling Technique) and bootstrapping have been widely adopted to balance class distributions and increase sample diversity [11]. However, these techniques often generate synthetic data without considering the semantic coherence of survey responses, which is critical in student satisfaction analysis. In contrast, Genetic Algorithm (GA)-based augmentation methods have gained traction for their ability to preserve the inherent structure and

logic of the original data [12]. GA simulates natural evolutionary processes, making it suitable for scenarios where maintaining the internal consistency of categorical and ordinal data is crucial.

Several researchers have integrated GA with predictive models to improve performance in educational analytics. For instance, a GA-based hybrid model was used to optimize input features in student performance prediction systems, leading to improved accuracy and interpretability [13]. The application of GA for data generation and feature selection has proven to be more effective in identifying critical variables in limited datasets. In studies comparing GA with other evolutionary algorithms like Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO), GA consistently outperformed in preserving data integrity and producing relevant synthetic samples for regression analysis [14].

Regression models such as LASSO (Least Absolute Shrinkage and Selection Operator) have been preferred for feature selection due to their ability to penalize insignificant variables and retain only the most impactful predictors [15]. LASSO has been widely used in educational research to identify and rank factors such as teaching quality, academic advising, campus facilities, and peer support. Its regularization strength helps in reducing multicollinearity and improving the robustness of the model, particularly when dealing with high-dimensional survey datasets.

Ordinal Regression, on the other hand, is especially suitable for modeling ordinal dependent variables such as satisfaction levels (e.g., highly satisfied, neutral, dissatisfied). Studies have demonstrated that Ordinal Regression provides better interpretability and predictive power in scenarios where the outcome variable has an inherent order but no uniform scale [16]. It has been applied to understand how various institutional factors influence the likelihood of students reporting different satisfaction levels.

In a comparative analysis of regression techniques for modeling student feedback, it was observed that combining data augmentation with LASSO and Ordinal Regression yielded improved prediction accuracy and more stable feature importance rankings [17]. Researchers highlighted that combining data augmentation with penalized regression techniques helps mitigate overfitting and improves generalization to unseen data. Moreover, the interpretability of the models makes them suitable for administrative decision-making in educational settings.

Additionally, several works have emphasized the importance of academic and career-oriented support in shaping student satisfaction. Empirical findings show that factors such as curriculum relevance, internship opportunities, mentorship programs, and placement support have a higher influence on satisfaction levels than infrastructure or recreational facilities [18]. These insights align with the present study's findings that career services, faculty guidance, and curriculum design play a more dominant role in determining student satisfaction.

Recent studies also propose the integration of artificial intelligence techniques, such as deep learning and ensemble learning, for predictive modeling in education. While these methods offer high accuracy, they often lack the transparency needed for policy formulation. Hence, there is a growing interest in combining interpretable models like LASSO and Ordinal Regression with evolutionary optimization techniques like GA for a balanced approach to prediction and explanation [19].

3. METHODOLOGY

3.1 DATA COLLECTION

The data for this study was collected through a well-structured survey. Through the survey various aspects of student satisfaction were captured. The survey is designed with questions including academic experiences, faculty support, learning resources, infrastructure facilities, extracurricular opportunities, administrative services, and career guidance and placement activities. The questions include demographic information including age, department and 36 other questions related to the mentioned areas. To ensure accessibility the survey was conducted through online mode using Google Forms, allowing students to participate from anywhere.

Students from different academic backgrounds and departments were invited to participate to ensure a diverse dataset. Before participating, students were informed about the study's purpose, and their consent was obtained. To encourage honest responses, the survey was designed to be anonymous, and students were assured that their feedback would be used solely for research purposes. The survey remained open for a sufficient period to allow students ample time to complete it. Through the survey 385 responses were collected

The collected data underwent an initial screening process to remove incomplete and irrelevant responses. This ensured that only high-quality and complete data was included in the final dataset for analysis. The systematic collection and screening of data helped in obtaining a reliable dataset that accurately represents student satisfaction levels.

3.2 DATA AUGMENTATION

The data set consists of the various factors through which the institution can understand students' satisfaction and dissatisfaction. It is experienced that collecting a very large amount of such data manually or online from students is a very cumbersome process. The earlier work shows that the larger the number of data points, the more accurate the result may be [10]. It is decided to enhance the dataset using machine learning techniques.

3.2.1 Genetic Algorithm:

Genetic Algorithm is the most appropriate method for data augmentation in student satisfaction analysis because it generates synthetic responses. It ensures that the augmented response will follow Likert scale structure. Traditional augmentation methods like random oversampling and SMOTE (Synthetic Minority Oversampling Technique), which either duplicate existing data or interpolate values inappropriately for categorical data. This makes it particularly useful for survey data, where responses must maintain an ordered structure [11]. Genetic Algorithm has the ability to reduce over fitting. Deep learning models need large data to augment but Genetic algorithms work well for small datasets and produce new data [12] Genetic Algorithm helps address class imbalance, a common issue in student satisfaction surveys where responses tend to be skewed towards positive ratings. Standard augmentation techniques may not preserve ordinal relationships, leading to inaccurate distributions. This improves regression model accuracy and feature selection,

leading to more reliable insights.GA-based augmentation enhances the robustness of student satisfaction analysis, prevents over fitting, and ensures that regression models capture meaningful patterns in survey responses. This makes it the most suitable technique for improving data-driven decision-making in education research.

GA is modeled after the Evolutionary Algorithm process, which is a subclass of GA. Often, genetic algorithms are utilized to produce superior solutions to and by depending on bio-inspired operators like selection, crossover, and mutation. The basic functionality of a Genetic Algorithm are selection, crossover and mutation

Genetic algorithms can be used for dataset augmentation by applying their key operators selection, crossover, and mutation to create new, diverse data samples [13]. Selection helps identify the most useful data points, crossover combines features from different samples to generate new ones, and mutation introduces small random changes to ensure variety. These techniques help enhance the dataset by adding realistic and diverse examples, improving the performance of machine learning models trained on the augmented data [14]. This section explains how these operators work and their role in generating high-quality augmented datasets.

- **Selection:** It is the process of picking individuals from the population to reproduce. Individuals with higher fitness scores are more likely to be chosen because they have better chances of producing good offspring.
- **Crossover:** It mimics biological reproduction by combining genetic material from two parents to create offspring. This helps mix good traits and explore better solutions. In Single-Point Crossover, the parent chromosomes are split at one point, and the parts are swapped. Two-Point Crossover uses two points for swapping, creating more variety. Uniform crossover [15] picks genes randomly from both parents for each position. The crossover process involves a random selection of genes to create a new gene from two parent genes. After copying the selected gene into a new chromosome, R1, another chromosome, R2, is chosen from a set of n chromosomes for crossover. Since performing crossover on genes is preferred, a random selection of certain genes, denoted as pi, is made from the newly generated gene set. Within these selected pi gene sets, random bits are chosen for crossover. The selected bits from the pi genes of R1 are then replaced. Genetic algorithms operate by generating and maintaining a population of entities represented as chromosomes.

Consider A chromosome R1 which has genes x1, x2, x3…, if one gene pi is selected randomly that contains the bits as shown in the Fig.1 below. Similarly, the same set of genes are selected from R2 which is shown in Fig.2.

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16

Fig.1. Bit patterns in Chromosome R1

Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	Y13	Y14	Y15	Y16

Fig.2. Bit patterns in Chromosome R2

Any random bits from the above given gene of R1 are selected, i.e. 3,7,9,12,15. Then these bits are replaced by the bits of the same position from Chromosome R2. Which will form a new bit pattern that is shown in Fig.3. This crossover gene is copied into a new chromosome.

X1	X2	Y3	X4	X5	X6	Y7	X8	Y9	X10	X11	Y12	X13	X14	Y15	Y16

Fig.3. Bit patterns of newly generated gene

- **Mutation:** Mutation is an essential operation in genetic algorithms that helps maintain genetic diversity in a population. It introduces small, random changes to the data. By modifying selected features in a dataset, mutation allows the algorithm to explore new possibilities and improve the quality of the generated data [16].

In this study, a dataset of student satisfaction responses is augmented using a genetic algorithm. The dataset consists of responses to survey questions, each rated on a five-point Likert scale, ranging from 1 (Very Dissatisfied) to 5 (Very Satisfied). The process of mutation creates realistic variants by carefully altering these responses. Since survey data consists of discrete categorical values, mutation must be applied in a way that maintains valid response values while introducing diversity. This is achieved using the rand mod method, which ensures randomness controlled.

To perform mutation, a random number must first be generated. This number serves as the basis for deciding how much the response value should change. The random number is generated using a pseudo random number generator (PRNG) or random number generator (RNG), which produces unpredictable values within a predefined range. In programming, functions such as randint(0, 100) generate a random integer between 0 and 100. Each time the function is called, a new number is selected. The random number allows for variation while ensuring that mutations remain structured. For example, if a random number of 27 is generated, it will be used in the next step to determine the mutation value. To calculate rand Mod for Mutation a random number (R) is generated, the modulus operation is applied to determine the mutation step. The modulus function calculates the remainder when R is divided by a predefined value, typically 5 (corresponding to the five response options in the Likert scale). This operation is represented as: if R = 27, then, 27 mod 5 will be calculated, that is the rand mod value will be 2.

The genetic algorithm is used for data augmentation. In this a gene is represented by the attribute of the dataset. The chromosome is a single data. In this case, each chromosome represents a specific set of augmentation techniques. The genes in these chromosomes can encode various facets of data augmentation. Each gene forces a different augmentation technique (e.g., rotation, scaling, noise injection). Or genes can also encode the parameters for each augmentation type. The values this gene can take are the options available in that gene.A set of numbers of chromosomes (data points) are initialized, called Initial Chromosomes.

In the fitness function a benchmark is designed and based on this a function that evaluates how well a generated chromosome fits the desired criteria. The criteria used in the fitness function of a genetic algorithm (GA) are necessary for guiding the evolutionary process towards ideal solutions. The fitness function

evaluates how well each individual (or chromosome) in the population solves the problem at hand, assigning a fitness score based on specific criteria that are relevant to the problem's objectives.

- To terminate the algorithm there is a criterion called *termination criteria,* which is the last point of the algorithm.
- The application of genetic algorithms (GA) is to increase the number of data sets. The hyper-parameter used in this algorithm is as follows
- Maximum number of iterations: This is termed as the maximum number of iterations; the algorithm will run if no other ending criteria are met. It sets an upper limit on the computational time and ensures that the genetic algorithm does not run forever.
- Population size: This states the number of individuals in each generation. A higher population might rises the diversity of solutions but also increase computational mandates.
- Mutation probability: This is the possibility that an individual solution will undergo mutation, which randomly alters one or more of its parameters. Mutation introduces variability into the population, possibly leading to determining new and better solutions.
- Elit ratio: The exclusiveness ratio determines the fraction of the top-performing individuals that are carried over to the next generation unchanged. This ensures that the best solutions are not lost during crossover and mutation.
- Crossover Probability: The probability with which two chromosomes will cross over parts of their solution to produce offspring. Crossover combines the characteristics of parent solutions, possibly creating more effective offspring.
- Parents Portion: Indicates what portion of the chromosome population will be selected as parents to produce the next generation. A higher portion means more of the chromosome population gets to reproduce, potentially rapid convergence but risking premature convergence to local optima.
- Crossover Type: Describes the method used for merging the solutions of two parents (e.g., uniform, one-point, two-point). The choice of crossover can affect how rapidly and effectively new solutions explore the search space.
- Maximum iterations without improvement: This sets an ending criterion based on a lack of improvement over a number of generations. If no improvement is witnessed for this many iterations, the algorithm terminates, assuming it has potentially converged.

These parameters are optimized to converge to a standard solution, with objective function as minimum as possible. In the case of this dataset the objective function is optimized as 16.05366659400288 as shown in the below graph Fig.(4) . This value helps as a measure of the "fitness" of the solution, which in this case assesses the deviation between the generated sample's statistical characteristics and those of the original dataset.

3.3 INDICATOR OF SOLUTION QUALITY:

In genetic algorithms, various solutions are created and tested, and each is assigned an objective function value. A value of 16.05366659400288 indicates that, for this particular solution, the cumulative difference in statistical properties from the original data is measured at approximately 16.05. This number helps in comparing different solutions.

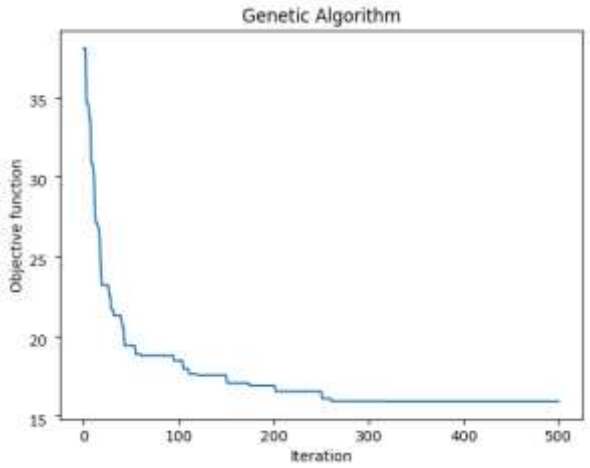

Fig.4. Graph of objective function v/s iteration

During the genetic algorithm's iterations, each new generation of results aims to achieve a lesser objective function value compared to previous generations. The reported value acts as a standard to gauge the effectiveness of the genetic algorithm in optimizing the parameters over following iterations. Essentially, this value is crucial for the genetic algorithm's process of evolutionary optimization, guiding the selection and breeding processes to hone in on the most effective solutions through comparative fitness assessment. This paper increases the dataset from 383 to 1000 data by introducing the above algorithm. The dataset is further evaluated using a regression method, which is discussed in the next section.

3.4 REGRESSION ANALYSIS

Regression analysis is a great statistical tool used to model relationships between a dependent variable and one or more independent variables. The aim is to understand how the distinctive value of the dependent variable changes when any one of the independent variables is changed, while the other independent variables are constant [17].

In the analyses conducted, ordinal regression shows effective for dependent variables that display ordered categories yet lack a consistent scale across responses. This modeling approach is mainly suitable for the data structure of this survey, wherein most responses to the questions are ordinal [18]. The application of ordinal regression facilitates:

1. The identification of key determinants driving satisfaction,
2. The prediction of varying levels of satisfaction,
3. The assessment of incremental effects on satisfaction.

Conversely, LASSO regression is engaged when dealing with a huge set of predictors. This method integrates variable choice with regularization, effectively sinking the coefficients of less

significant variables to zero, thereby improving model easiness and interpretability [19]. LASSO regression is employed to

- Select appropriate features by eliminating redundant predictors,
- Develop predictive models that are strong and efficient,
- Address issues of multicollinearity among predictors,
- Scale efficiently to complex survey data structures.

The ordinal regression models the ordinal nature of satisfaction responses, enabling an elegant understanding of how various factors successively affect satisfaction levels, progressing through categories such as 'Neutral' to 'Satisfied'. LASSO regression, on the other hand, focuses on separating the most influential predictors, thereby constructing a efficient model that emphasizes the primary drivers of satisfaction in datasets characterized by extensive predictor variables. The following section delves into the theoretical foundations of ordinal regression.

3.5 ORDINAL REGRESSION

The most common ordinal regression model is the Proportional Odds Model (POM) or Cumulative Logit Model. This works as follows:

It calculates the cumulative probabilities for each category:

$$P(Y \le k) = P(Y = 1) + P(Y = 2) + \cdots + P(Y = k) \quad (1)$$

where k is the category of interest, P represents the probability function. It assigns a probability to each outcome in the sample space of the random variable Y.

3.5.1 Log Odds Transformation:

It applies a logit transformation to the cumulative probabilities' logit:

$$\text{logit}(P(Y \le k)) = \ln\left(\frac{P(Y \le k)}{1 - P(Y \le k)}\right) \quad (2)$$

The relationship is modeled as logit is expressed as in Eq.(3)

$$\text{logit}(P(Y \le k)) = \theta_k - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_p X_p \quad (3)$$

where θ_k is threshold or intercept for category k and β is the Coefficients of the predictors ($X_1, X_2, \ldots, X_p$).

In the Proportional Odds Assumptions, assumes that the effect of predictors is consistent across all thresholds. For example, the influence of a factor like "teaching quality" on moving from "Neutral" to "Satisfied" is the same as moving from "Satisfied" to "Very Satisfied." Steps to Perform Ordinal Regression:

1) Define the Variables:
   a) Ensure the dependent variable is ordinal.
   b) Identify predictors (independent variables) that could influence the dependent variable.
2) Model Fitting:
   a) Use statistical software to fit the ordinal regression model.
3) Check Assumptions:
   a) Proportional Odds: Test if the relationship between predictors and log odds is consistent across categories.
   b) If the assumption is violated, consider alternative models like generalized ordinal regression.
4) Interpret Results:
   a) Coefficients (β): Show how predictors influence the odds of being in higher versus lower categories.
   b) Thresholds (θ_k): Define the cutoff points between categories.

3.6 LASSO REGRESSION

The full-length of LASSO regression is given as Least Absolute Shrinkage and Selection Operator. This is a type of linear regression. This also adds a price or regularization to the model to prevent over fitting and improves prediction accuracy [20]. It is particularly effective when dealing with datasets that have many forecasters, particularly if some of them are trivial.

LASSO regression modifies the standard linear regression objective function by adding a penalty term proportional to the sum of the absolute values of the regression coefficients. This penalty has two key effects

- Coefficients are forced to be smaller, indicating Shrinkage
- Some coefficients are reduced to exactly zero, effectively removing irrelevant predictors from the model, called as Variable Selection

The LASSO objective function is to:

$$\text{Minimize} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}|\ \beta_j| \quad (4)$$

where $\frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of square or ordinary least square term. λ is Regularization parameter controlling the strength of the penalty. The $\sum_{j=1}^{p}|\ \beta_j|$ is L1 norm penalty (sum of absolute values of coefficients).

This LASSO regression will shrink some coefficients to exactly zero, effectively removing non-important predictors from the model. This makes it useful for datasets with many predictors, as it automated variable selection. This also helps prevent overfitting by penalizing large coefficients. Reducing the number of predictors creates simpler, more interpretable models. It will find that only a subset of predictors will have non-zero coefficients.

3.6.1 Steps in LASSO Regression:

1) **Data Preparation**:
   a) Standardize predictors to ensure coefficients are comparable (since LASSO depends on the magnitude of coefficients).
2) **Select λ**:
   a) The regularization parameter λ controls the strength of the penalty:
      i) If λ=0, LASSO becomes standard linear regression.
      ii) As λ increases, more coefficients shrink to zero.
   b) Use techniques like cross-validation to find the optimal λ.
3) **Model Fitting**:

a) Fit the LASSO regression model using software tools

4) **Interpret Results**:

a) Identify significant predictors with non-zero coefficients.

b) Use the model for prediction or insight into variable importance.

LASSO regression is a powerful tool for Feature selection in high-dimensional datasets. Creating interpretable and robust predictive models. Addressing overfitting by controlling model complexity.

By focusing on the most influential predictors and ignoring the rest, LASSO ensures the final model is both practical and accurate, making it highly applicable in fields like survey analysis, genomics, marketing, and more. Train a deep learning model. It is not sufficient that only the factor influencing the output but also it is required to find how and why it influences hence we use explainable AI for the next processes Even if you evaluate in deep learning model we will not understand why it happens to understand that the results of explainable AI is required.

4. RESULT AND DISCUSSION

4.1 INTERPRETATION OF HEATMAP (FEATURE CORRELATIONS)

The Fig.5 illustrates the correlation between various student satisfaction factors.

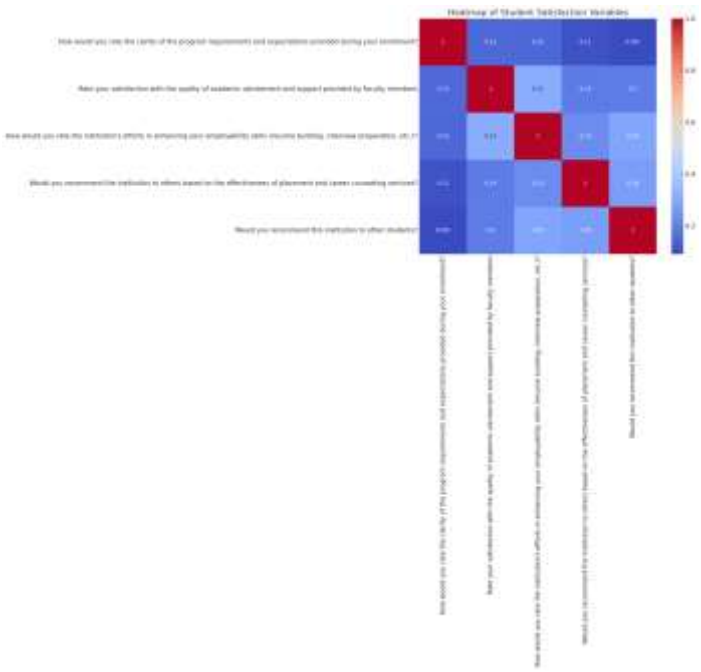

Fig.5. Heatmap of student satisfaction variables

The colour intensity shows the strength of relationship among factors. The strong positive correlations indicated by the red colour, an increase in one factor leads to an increase in another factor. Here career services and curriculum are strongly correlated with overall satisfaction. Which recommends that improving in these areas will significantly enhance students' satisfaction towards the institution. Blue shades represent negative correlations, where an increase in one factor leads to a decrease in another, such as ineffective administrative communication reducing student trust and satisfaction. Lighter shades indicate weak or no correlation, meaning changes in one factor do not significantly influence another. Features like Wi-Fi quality and library facilities, while appreciated, do not directly impact overall satisfaction. Clarity of program requirements is strongly correlated with satisfaction in academic advice, suggesting that students who clearly understand their course structure are more likely to be satisfied with faculty support. Employability skills enhancement is highly correlated with placement service effectiveness, which shows that if institutions can provide career and placement training it will lead to better placement for the students. The overall recommendation of the institution is closely related to satisfaction with program advisement, clarity of the program, and placement training services. Library facilities and administrative communication show weak correlations with overall satisfaction. By analyzing these correlations, institutions can improve areas like career counseling, faculty mentorship, and industry-aligned curricula.

4.2 LASSO REGRESSION HEATMAP (FEATURE IMPORTANCE)

LASSO regression was applied to perform feature selection. By identifying the most influential factors that contribute to student satisfaction by shrinking less important coefficients to zero. The heatmap in Fig.visually represents these selected features, where only variables with significant predictive power retain non-zero coefficients. Among the most impactful factors, Placement & Career Counseling Services (0.1839) emerged as the strongest predictor, indicating that students who receive effective career guidance and job placement support are more likely to be satisfied and recommend their institution. Similarly, an Industry-Relevant Curriculum (0.1519) was found to play a crucial role, as students prefer academic programs that align with real-world job market needs, enhancing their confidence in future employability. Another key factor, Faculty Support & Academic Advisement (0.1041), emphasizes the importance of faculty mentorship in shaping a positive academic experience. When faculty members provide strong academic guidance, students are more engaged and satisfied with their learning journey. Additionally, Extracurricular Activities & Institutional Events (0.1002) contribute significantly to satisfaction, as participation in co-curricular programs enhances students' skill development, networking opportunities, and overall engagement. Conversely, factors such as Wi-Fi availability, administrative communication, and HVAC systems were found to have insignificant coefficients, meaning they had little to no impact on overall satisfaction. This suggests that while these facilities contribute to convenience, they do not play a defining role in shaping students' overall perception of their educational experience. These findings reinforce the need for institutions to prioritize academic quality, career support, and faculty engagement over less influential infrastructural improvements when aiming to enhance student satisfaction.

4.3 ORDINAL REGRESSION HEATMAP (MODEL COEFFICIENTS)

The Fig.7 presents how various factors influence student's satisfaction with Ordinal Regression heatmap.

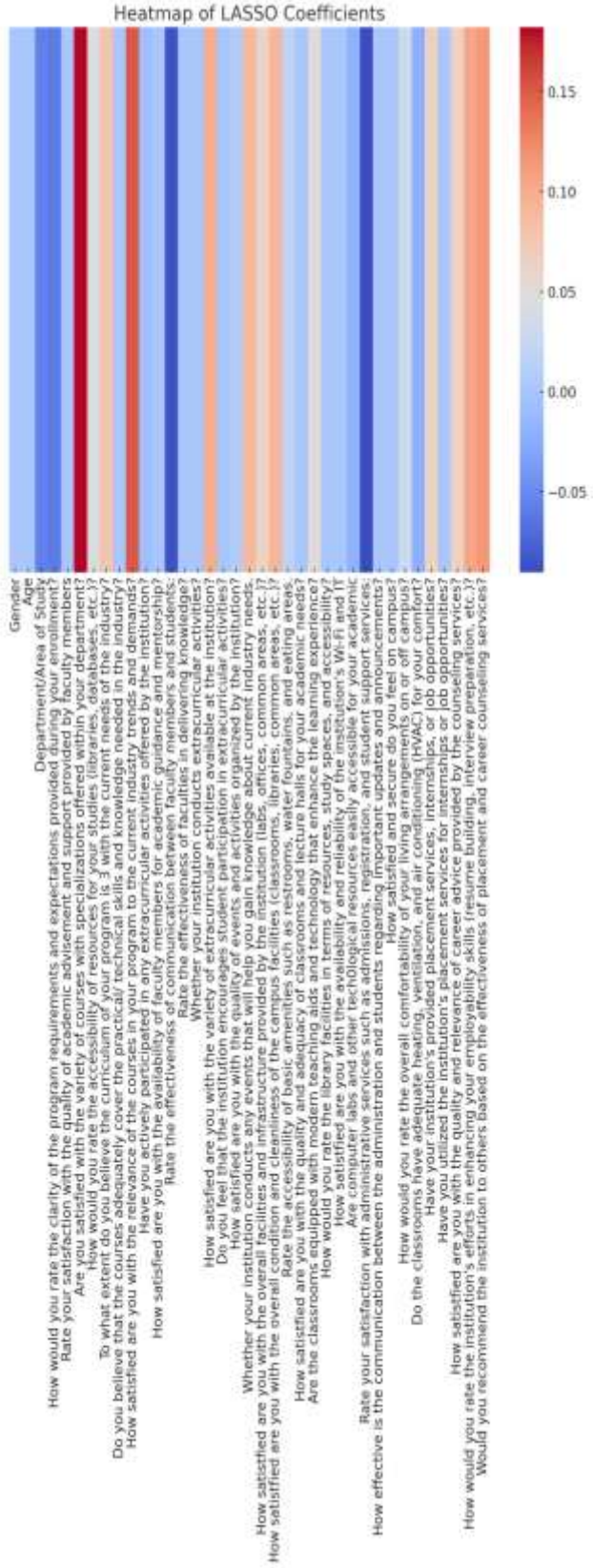

Fig.6. Heatmap of Lasso Regression

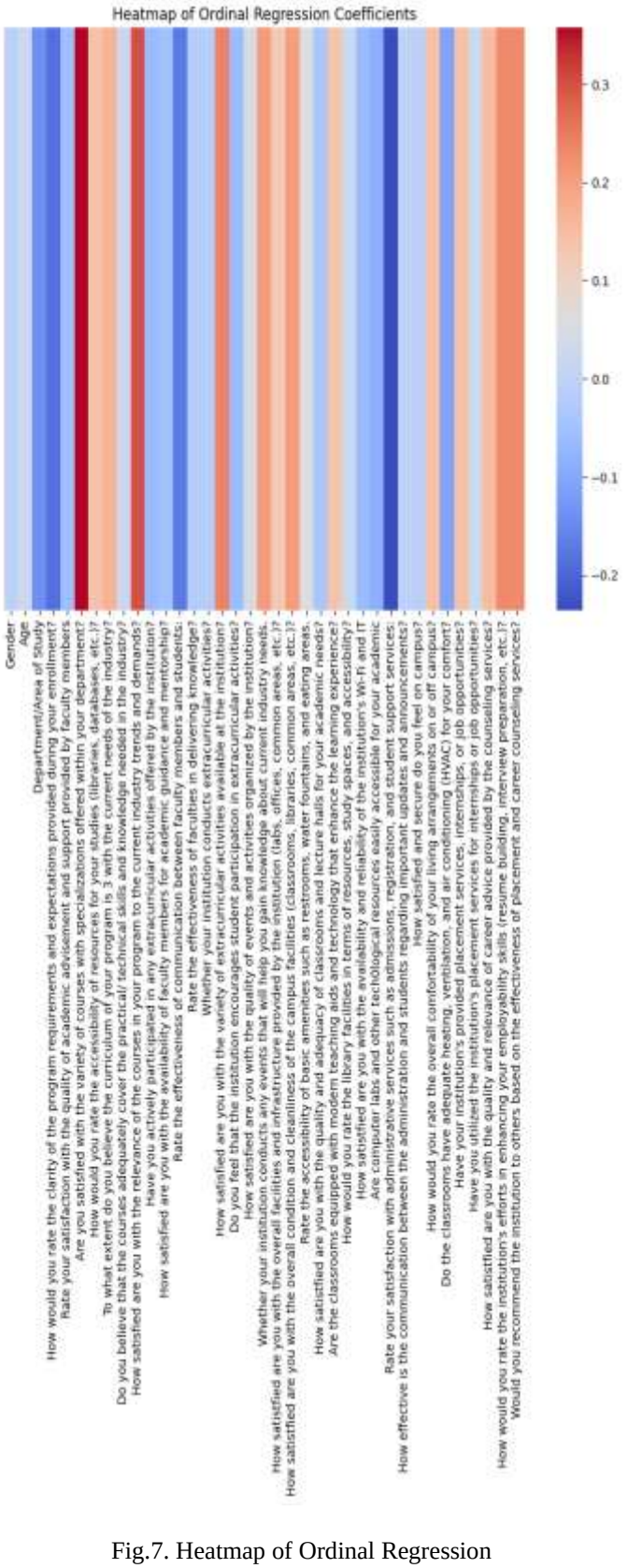

Fig.7. Heatmap of Ordinal Regression

Ordinal Regression is well suited for Likert-scale data. Which calculates the likelihood of students selecting higher or lower ratings from the options. The heatmap highlights positive coefficients for Placement Activities, Curriculum Design, and

Faculty Support. The analysis shows that career guidance, industry-aligned syllabus, and faculty support significantly enhance student satisfaction. Placement services play a crucial role, which has the main role in shaping student perception. Conversely, negative coefficients indicate that administrative inefficiencies and poor communication lower satisfaction. Poor, unclear rules and delayed responses will reduce the student's trust. Gender and Wi-Fi quality show no significant impact on satisfaction. These findings suggest that institutions should prioritize faculty advice, curriculum relevance, and career counseling over secondary concerns like amenities, ensuring a stronger impact on student satisfaction and institutional reputation.

The Genetic Algorithm (GA) was fine-tuned to generate realistic student satisfaction survey responses while maintaining the ordered structure of Likert-scale data. The objective function, reduced to 16.05366659400288 (Fig.4), ensures that the synthetic dataset reflects the characteristics of actual survey responses, making it suitable for Ordinal Regression, which predicts the likelihood of students selecting different satisfaction levels. Unlike LASSO Regression, which emphasizes feature selection by eliminating less significant variables, Ordinal Regression retains the ranking structure of responses, offering a clearer picture of how different factors shape student satisfaction. The results indicate that Placement Services (0.1839), Industry-Relevant Curriculum (0.1519), Faculty Support (0.1041), and Extracurricular Activities (0.1002) have a strong positive effect on satisfaction, increasing the chances of students providing higher ratings. On the other hand, Administrative Inefficiencies (-0.0927) and Ineffective Communication (-0.0863) negatively impact satisfaction, highlighting the role of institutional management in shaping student perceptions. Additionally, factors like Gender and Wi-Fi availability showed little to no influence, suggesting that they are not key determinants of overall satisfaction. By applying GA-based data augmentation with Ordinal Regression, this study presents a structured approach that preserves the ordinal nature of survey responses while improving analysis accuracy. The findings confirm that academic and career-related factors play a crucial role in student satisfaction, whereas infrastructure-related concerns have a lesser impact. To enhance student experiences and strengthen institutional reputation, universities should focus on career counseling, curriculum development, and faculty mentorship, along with better administrative communication.

5. CONCLUSION

In this study, data augmentation using Genetic Algorithm enhances regression analysis by increasing the data, providing better insights into student satisfaction. Using Regression analysis, it is evaluated that quality of academic activities, placement support, and faculty engagement are the important factors of satisfaction. Educational institutions should focus more on areas of secondary concerns like infrastructure and Wi-Fi quality. Future research for this study is to incorporate machine learning models and explainable AI techniques to extract predictions and improve institutional decision-making.

REFERENCES

[1] E. Razinkina, L. Pankova, I. Trostinskaya, E. Pozdeeva, L. Evseeva and A. Tanova, "Student Satisfaction as an Element of Education Quality Monitoring in Innovative Higher Education Institution", *E3S Web of Conferences*, Vol. 33, pp. 1-6, 2018.

[2] M. Kuzehgar and A. Sorourkhah, "Factors Affecting Student Satisfaction and Dissatisfaction in a Higher Education Institute", *Systemic Analytics*, Vol. 2, No. 1, pp. 1-13, 2024.

[3] N. Zakaria, R. Umar, W.H.A.W. Deraman and S.S.S.A. Mutalib, "Regression Analysis on Factors Influencing Students' Satisfaction towards Programme Courses", *Indian Journal of Science and Technology*, Vol. 9, No. 17, pp. 1-5, 2016.

[4] B. Kumar and R. Kumar, "Generalizing Clustering Inferences with ML Augmentation of Ordinal Survey Data", *Computer Science*, Vol. 25, No. 1, pp. 1-8, 2024.

[5] M.A. Lateh, A.K. Muda, Z.I.M. Yusof, N.A. Muda and M.S. Azmi, "Handling a Small Dataset Problem in Prediction Model by Employ Artificial Data Generation Approach: A Review", *Journal of Physics: Conference Series*, Vol. 892, No. 1, pp. 2-8, 2017.

[6] Y. Chen, M. Elliot and J. Sakshaug, "A Genetic Algorithm Approach to Synthetic Data Production", *Proceedings of International Workshop on AI for Privacy and Security*, pp. 1-4, 2016.

[7] W. Chen, "Application of Immune Genetic Algorithm in Survey Data Processing", *Applied Mechanics and Materials*, Vol. 241, pp. 1737-1740, 2013.

[8] U. Mehboob, J. Qadir, S. Ali and A. Vasilakos, "Genetic Algorithms in Wireless Networking: Techniques, Applications and Issues", *Soft Computing*, Vol. 20, pp. 2467-2501, 2016.

[9] D.E. Goldberg and J.H. Holland, "Genetic Algorithms and Machine Learning", *Machine Learning*, Vol. 3, No. 2, pp. 95-99, 1988.

[10] P.S. Nethravathi and K. Karibasappa, "Business Intelligence Appraisal of the Customer Dataset based on Weighted Correlation Index", *International Journal of Emerging Technology and Research*, Vol. 3, No. 6, pp. 31-41, 2016.

[11] H. Mansourifar and W. Shi, "Deep Synthetic Minority Over-Sampling Technique", *Proceedings of International Conference on Machine learning*, pp. 1-9, 2020.

[12] C.M. Villegas, J.L. Curinao, D.C. Aqueveque, J. Guerrero-Henríquez and M.V. Matamala, "Data Augmentation and Hierarchical Classification to Support the Diagnosis of Neuropathies based on Time Series Analysis", *Biomedical Signal Processing and Control*, Vol. 95, pp. 1-6, 2024.

[13] Mehboob, Junaid Qadir, Salman Ali and Athanasios Vasilakos, "Genetic Algorithms in Wireless Networking: Techniques, Applications and Issues", *Soft Computing*, Vol. 20, No. 6, pp. 1-5, 2016.

[14] Lawerence Davis, "Handbook of Genetic Algorithms", 1991.

[15] K. Deep and M. Thakur, "A New Crossover Operator for Real Coded Genetic Algorithms", *Applied Mathematics and Computation*, Vol. 188, No. 1, pp. 895-911, 2007.

[16] S. Ullah, A. Salam and M. Masood, "Analysis and Comparison of a Proposed Mutation Operator and its Effects on the Performance of Genetic Algorithm", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 25, No. 2, pp. 1208-1216, 2022.

[17] M. Sarstedt, E. Mooi, M. Sarstedt and E. Mooi, "Introduction to Market Research", *A Concise Guide to Market Research: The Process, Data and Methods using IBM SPSS Statistics*, 209-256, 2019.

[18] M. Lalla, "Fundamental Characteristics and Statistical Analysis of Ordinal Variables: A Review", *Quality and Quantity*, Vol. 51, pp. 435-458, 2017.

[19] J. Ranstam and J.A. Cook, "LASSO Regression", *Journal of British Surgery*, Vol. 105, No. 10, pp. 1-5, 2018.

[20] L. Zhang, X. Wei, J. Lu and J. Pan, "Lasso Regression: From Explanation to Prediction", *Advances in Psychological Science*, Vol. 28, No. 10, pp. 1-7, 2020.