

University Information System's Impact on Academic Performance: A Comprehensive Logistic Regression Analysis with Principal Component Analysis and Performance Metrics

Aybeyan Selim¹, Ilker Ali^{1,2}, Blagoj Ristevski²

¹ Faculty of Engineering and Architecture, International Vision University, Gostivar, North Macedonia

² Faculty of Information and Communication Technologies – Bitola, University "St. Kliment Ohridski" – Bitola, North Macedonia

Abstract – This paper comprehensively analyzes data mining techniques and performance metrics applied to a logistic regression model. Principal Component Analysis (PCA) was utilized to diminish the complexity of high-dimensional data, enabling clearer visualization and examination of intricate relationships among variables. The logistic regression model demonstrated commendable performance on both test and train sets, as evidenced by high values of accuracy, precision, recall, ROC AUC, and F1 Score were observed. The provided confusion matrices offered detailed insights into the model's accuracy in classifying positive and negative instances. Concerning our hypotheses, we found no significant relationship between gender and academic performance, supported by a highly significant p-value of 0.0 and a weak positive correlation coefficient of 0.0847. However, we noticed a strong positive correlation of 0.99 between gender and exam characteristics, although it has not reached statistical significance for a p-value of 0.281. Our research contributes valuable insights into data analysis, model evaluation, and the interplay between variables.

DOI: 10.18421/TEM132-72

<https://doi.org/10.18421/TEM132-72>


Corresponding author: Aybeyan Selim,
Faculty of Engineering and Architecture, International
Vision University, Gostivar, North Macedonia
Email: aybeyan@vision.edu.mk

Received: 19 November 2023.

Revised: 03 March 2024.

Accepted: 11 March 2024.

Published: 28 May 2024.

 © 2024 Aybeyan Selim, Ilker Ali & Blagoj Ristevski; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

The findings can inform decision-making in real-world applications and warrant further investigation of identified relationships to enhance practical implications. Future studies should consider exploring additional factors, such as the subject of study, semester, and study year, to further understand their impact on student performance.

Keywords – Principal component analysis (PCA), logistic regression, data analysis, machine learning and performance metrics.

1. Introduction

A system is defined as a set of objects between which there is a certain connection and exchange of information and through which it realizes a certain function [1]. Thus, the system represents a functional set of objects and links between them. Objects and their interrelationships are described by properties called features or attributes. Systems can be diverse that depends on the things that make up the system and the function that the system performs [2].

One extremely complex system is the university. Such complex systems represent a set of subsystems, each with a specific function. The university is one system, while the faculties represent a separate subsystem within that system [3]. The system size depends on the number and size of the objects it contains. Each system is separated from the outside, which is called the environment of the system. The system can be completely closed about its environment. On the other hand, most systems communicate with the environment.

The technical and personnel specifications of the information system are precisely defined. The technical specification refers to the scope of the system, required equipment, and software for managing processes within the system, the size and number of implemented databases, and the like.

The personnel specification includes the number of trained personnel required to handle, maintain, and use the information system [4].

The information system is viewed as a processing process based on input-generating output data [5]. Bringing input data and taking output data from the system is achieved using "data flows". Data flows in the university information system (UIS) are grades, teaching materials, printed documents, messages, electronic documents, etc. [6]. User interfaces are web and mobile data sources and sinks. These objects outside the information system communicate with the system by sending or receiving information.

Data stores are deferred or accumulated streams of data. Users' demographic data includes various records, archives, or files [7]. Data flow diagrams show the interface or storage connection as data sources/sinks, with corresponding processes, and the interconnection of processes. The diagram at the highest level of abstraction represents the information system as one approach that communicates with the environment through interfaces and data flows [8]. Lower-level diagrams are obtained by decomposing the highest-level diagram. Decomposition is the breaking down of the basic process into parts, i.e., sub-processes.

1.1. University Information System at International Vision University

The Information System at International Vision University (IVU) serves as a vital platform for facilitating coherent connections between users through information exchange [6]. The UIS plays a pivotal role in supporting the higher education process. As a supporting system, UIS is more comprehensive and is developed to support the education process and close the weak points in the University. Academic Information System (AIS), Student Service (SS), and the university's database are interweaved within UIS, ensuring a cohesive and comprehensive system, shown in Figure 1. With its user-friendly approach and robust security measures, UIS remains a valuable asset in the university's pursuit of success. The general features of UIS are:

Data/Log Processing: UIS efficiently handles record storage and various data processing functions.

Integrated Database: Utilizing an integrated database, UIS supports and extends functionalities across different areas.

Access to Timely Information: UIS empowers operational, tactical, and strategic level managers with easy and timely access to essential information.

Flexibility: Designed to adapt to the University's evolving needs, UIS remains flexible and responsive.

Security: UIS maintains strict access control ensuring only authorized persons can access sensitive information.

User-Friendly Accessibility: UIS is designed for ease of use, ensuring effortless accessibility to authorized personnel.

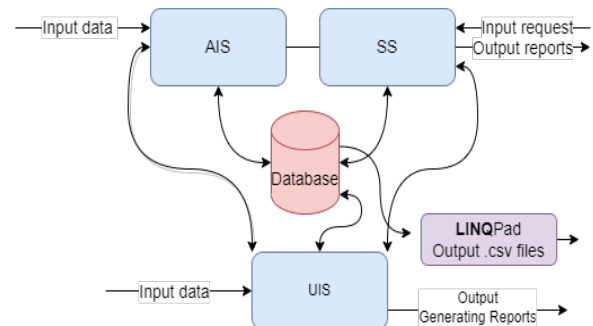


Figure 1. Information system architecture at International Vision University

2. Related Works

Information and Communication Technology (ICT) systems have a pivotal role in improving the overall student learning experience within higher education institutions. These systems effectively capture data from various sources, thereby improving relevance, efficiency, and overall effectiveness in educational settings [9]. In a parallel vein, the implementation of Outcome Based Education (OBE) information systems has demonstrated significant promise in measuring and improving student learning outcomes. Safiudin *et al.* emphasize the positive impact of such systems on academic quality and student achievement, particularly in college study programs [10]. The development of web-based OBE information systems signifies a step forward in aligning educational practices with desired learning outcomes [11].

Academic analytics, utilizing student data, emerges as a potent tool for predicting academic performance and identifying crucial predictor variables. The modeling of engineering student academic performance through academic analytics provides insights into the predictive capabilities of this methodology [12].

Innovative pedagogical approaches, such as those implemented by Unicorn College in Prague, Czech Republic, demonstrate the transformative potential of interactive technologies. Beranek, Bory, and Vacek [13] outline the utilization of interactive textbooks, podcasts, and Nearpod software applications to support student learning, demonstrating a commitment to improving learning outcomes through diverse and engaging tools.

Holderied [14] explores the intersection of interactive technologies and active learning strategies, specifically in the context of information literacy classes. The integration of these methods has been demonstrated to boost student engagement and, consequently, enhance learning outcomes. The research emphasizes the importance of dynamic instructional design in fostering effective library instruction.

Data mining applications in educational settings emerge as a powerful mechanism for improving the student learning experience. Abdous, He, and Yen [15] illustrate how data mining techniques can analyze learning patterns, interactions, and live video streaming environments to provide valuable insights. Such insights enhance the learning experience's overall quality and refine educational strategies.

eCamp, a visual knowledge discovery system, adds a novel dimension to student progression analysis. Raji *et al.* [16] introduce a system that reveals previously unavailable information, aiding in student outcomes, retention efforts, and curriculum design. The visual representation of student records data proves instrumental in offering a comprehensive understanding of student progress.

Acknowledging the heterogeneity among students in constructing predictive models emerges as a crucial consideration for improving performance and identifying vulnerable students more precisely. Helal *et al.* advocate for the incorporation of rule-based and tree-based methods in predictive modeling, emphasizing the interpretability and effectiveness of such models in predicting academic performance [17].

Researchers extensively explore student attitudes, learning experiences, and success enhancement across educational settings [18], [19]. Harrell and Bower's study [20] identifies key factors for student success in online courses, guiding strategies crucial for engagement, retention, and degree completion. Address challenges in college readiness for minority students in mathematics, offering insights for targeted interventions [21].

In higher education, Hu, McCormick, and Gonyea's research [22] on student learning and persistence contributes significantly to understanding factors supporting students' academic journey. Digital technologies in distance learning [23], [24], effective online learning strategies [25], and comparative analyses [26] enhance outcomes and forecast student success.

Studies on factors influencing students' preference for online learning [25] and assessing open-source learning management systems [27] provide valuable insights. Student satisfaction in transnational higher education informs practices for international settings [28].

Digital technologies significantly impact student learning experiences and academic achievement [29]. Cultivating digital competencies is crucial for student success [30], contributing to ongoing efforts for academic excellence.

The importance of academic information systems is stressed, supported by Mbaeze, Ukwandu, and Anudu's study [29] linking ICT adoption to academic performance. Wu, Lin, and Tsai's [31] predictive model aligns with the broader context of improving academic performance. The study on Finnish eighth graders' aspirations [32] and the examination of digital technology risks [33] contribute to understanding factors influencing educational outcomes, aligning with Fobel and Kolleck's [34] study on cultural education disparities in Germany.

3. Data and Methodology

In this section, we delineate the core components of our study, offering a clear overview of our research objectives, hypotheses, algorithms utilized, evaluation metrics, research methodology, and the dataset employed for analysis.

Aim of the study

This study aims to leverage machine learning algorithms to analyze and assess data from the UIS, extracting valuable insights and enabling data-driven decision-making to enhance various facets of the university's operations and student outcomes. By scrutinizing data from the UIS through machine learning algorithms, this study harnesses the potential of data analytics to propel evidence-based decision-making, augment student success, and optimize diverse facets of university operations.

Hypothesis

This research study examines the correlation between the UIS and students' academic performance and behavior. We hypothesize that UIS plays a significant role in shaping students' performance and behavior, particularly concerning their gender. Through the utilization of machine learning algorithms, our objective is to extract valuable insights into the influence of UIS on students' academic outcomes and behavioral patterns.

Hypothesis 1:

Our first hypothesis aims to investigate the association between students' academic performance and their gender.

H₀: There exists a noteworthy correlation between academic performance and gender.

H_a: There is no notable correlation between academic performance and gender.

Utilizing the comprehensive dataset from UIS, we will conduct statistical tests, including Chi-square and Pearson correlation, to explore the potential link between UIS usage and students' academic achievements.

By making this analysis, we seek to identify whether UIS usage has a discernible influence on students' academic success and whether gender plays a significant role in this relationship.

Hypothesis 2:

Our second hypothesis delves into understanding distinct behavior patterns among students based on various exam characteristics, such as colloquium, student status (full-time/part-time), seminary attendance, engagement in other activities, final exams, and make-up exams.

H₀: There is notable correlation between exam characteristics and gender.

H_a: There is no notable correlation between exam characteristics and gender.

Additionally, we investigate how gender interacts with these exam characteristics. Employing advanced machine learning algorithms on the UIS data, we aim to reveal meaningful insights into how UIS usage may impact students' behavior patterns differently across genders.

Algorithms

In the study, principal component analysis and logistic regression machine learning algorithms were used, and the details of these algorithms are given in the following two sections.

3.1. Principal Component Analysis

Principal component analysis (PCA) is a robust and commonly employed statistical method that is fundamental to data analysis and dimensionality reduction. At its core, PCA seeks to reduce the dimensionality of a high-dimensional dataset into a lower-dimensional space, thereby preserving the maximum variance in the original data. By doing so, PCA facilitates the representation of complex datasets in a more manageable and interpretable manner, enabling efficient computation and insightful visualization. The main goal of PCA is to identify principal components, which are orthogonal vectors that capture the directions of maximum variance within the data [35]. Through this process, PCA enables the reduction of the dataset's dimensionality by selecting a subset of the most informative principal components while still preserving a significant portion of the original data's variance. Consequently, PCA aids in simplifying complex datasets, making them amenable to further analysis and interpretation. PCA involves projecting the features onto a reduced representation. Given a training set comprising n training examples denoted as $X = \{x_1, x_2, \dots, x_n\}$, PCA generates principal components P_k that serve as linear combinations of the original features X [36]. This can be written as

$$P_k = a_{k_1}x_1 + a_{k_2}x_2 + \dots + a_{k_n}x_n \quad (1)$$

where $\sum_{i=1}^n a_{k_i}^2 = 1$

3.2. Logistic Regression

Logistic Regression is a fundamental and extensively employed statistical technique that holds a prominent place in the field of predictive modeling and binary classification [37]. Logistic regression primary objective is to establish a model that characterizes the relationship between a binary variable and one or more independent predictors, often referred to as predictors or features. The dependent variable, typically represented as a binary outcome, is transformed using the logistic function, which maps continuous values into the range of [0,1]. This transformation allows logistic regression to estimate the probability of an event occurring, making it well-suited for binary classification tasks.

The sigmoid function also known as a logistic function, is defined as follows:

$$P(y = 1) = \frac{1}{1 + e^{-z}} \quad (2)$$

where $P(y = 1)$ is probability of the positive outcome (e.g., class 1), and z is a linear combination of independent variables and their respective coefficients:

$$z = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n, \quad (3)$$

Here, $\beta_0, \beta_2, \dots, \beta_n$ are the coefficients, also known as weights, estimated through a process called maximum likelihood estimation, aiming to discover the optimal model that maximizes the observed data probability.

The logistic regression model allows for the interpretation of the relationship between the independent variables and the probability of a positive outcome. By analyzing the coefficients, one can determine the direction and strength of the impact each predictor has on the outcome. Additionally, logistic regression provides a valuable tool for understanding the statistical significance of each predictor, aiding in feature selection and model refinement [38].

3.3. Evaluation Metrics

The performance of each model has been assessed using numerous metrics, such as accuracy, precision, recall, and the F1 score, supplemented by the utilization of the confusion matrix [39].

The confusion matrix serves as an essential instrument for verifying the performance of a classification model. It is a 2x2 matrix representing the true and predicted classes. In this study, since we are working with a dataset consisting of two classes, the confusion matrix takes the form shown in Table 1.

Table 1. Confusion matrix

Confusion Matrix		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Accuracy: The metric measures of the accuracy measure the ratio of correct predictions (True Positives (TP) and True Negatives (TN)) to the total samples evaluated (TP, TN, False Positives (FP), and False Negatives (FN)). The formula for accuracy is represented as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision: The ratio of TP to the sum of TP and FP is precision. It indicates the model's capacity to recognize positive instances accurately.

$$Precision = \frac{TP}{FP + TP} \quad (5)$$

Recall: The model's sensitivity or true positive rate is the ratio of TP to the sum of TP, and FN is called recall. It assesses the model's capability to capture all positive instances.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

F1 Score: Precision and recall harmonic mean represents the F1 score of the model. It offers a balanced measure between precision and recall, proving useful particularly when there is an uneven class distribution.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

3.4. Data Set

The data set comprises demographic information, academic performance, exam characteristics, department, semester, subject code and realized traffic of students in UIS user interface with a total of 27,271 individuals. Of these, 10,801 are identified as female, while 17,470 are identified as male. The detail of the dataset by attribute gender and grade is given in Table 3.

Table 2. Details of data set

	5 (.< 51)	6 (51-60)	7 (61-70)	8 (71-80)	9 (81-90)	10 (91-100)
Male	5032	1898	1854	3129	1982	3574
Female	2488	1163	1279	1886	1379	2606

The data collection process includes two main steps. Firstly, the administrator extracts data directly from the MS-SQL server in the form of a .csv file.

Secondly, data related to student activities and traffic on UIS are downloaded from Microsoft Portal 365. To consolidate and correlate the information, the admin uses the email of each student as the primary key, facilitating the joining of the two datasets from MS-SQL and Microsoft Portal 365. This approach ensures a comprehensive and unified dataset for further analysis and insights.

4. Methodology

This study tested a dataset from UIS for identifying students' academic performance, eventual graduation rates, and exam characteristics to the attribute gender. The dataset has been imported into Python, followed by preprocessing steps and the modeling stage applied to the dataset. Figure 2 presents the followed methodology of this study in summary.

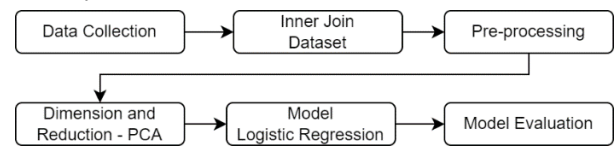


Figure 2. Methodology

In this study, all the processes depicted in Figure 2 were implemented in Python 3.10.1 programming language. Various libraries were utilized for different operations with the dataset. Specifically, NumPy, Pandas, and scikit-learn libraries were employed for data manipulation and machine learning tasks. Additionally, Matplotlib and Seaborn libraries were employed to visualize the results of the models. The codes were executed using the Jupyter Notebook platform, providing an interactive and collaborative environment for data analysis and code execution. The libraries used in Python for each process are summarized in Table 2.

Table 3. Python libraries used in the research

Python Library	Process	Phase
Glob	Dataset reading	Data
NLTK	Pre-processing and downloading of stop words	Processing
re	Numbers punctuation and discarding	
NumPy	Matrix operations	Modelling
Pandas	Data manipulation and analysis	
scikit-learn	Execution of machine learning algorithms	
matplotlib	Visualization	Evaluation
Seaborn	Visualization	of Modeling
SciPy	Statistics	

5. Experimental Results and Analysis

PCA is a fundamental technique in data analysis and machine learning. The "Before PCA" section of the figure illustrates the original high-dimensional data representation. In its raw form, the data may contain multiple features, making visualization and analysis challenging due to the complex interdependencies between variables. This high-dimensional space could lead to difficulties in capturing the most significant patterns and could potentially increase computation time for machine learning algorithms.

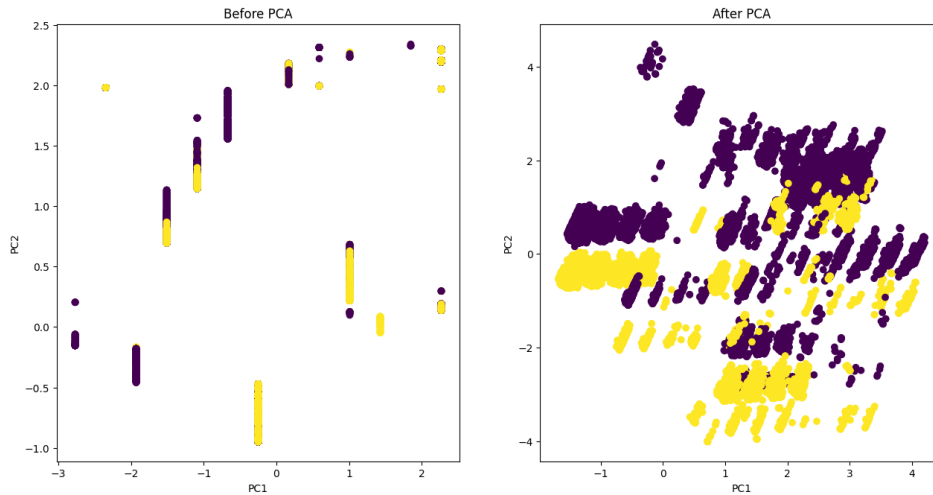


Figure 3. Dimension and Reduction with PCA

Accuracy, precision, recall, ROC AUC, and F1 Score are vital performance metrics utilized to assess the efficacy of classification models. In this article segment, we present a comprehensive analysis of these metrics for a logistic regression model on both the test set as well as the training set. The outcomes are shown in Table 4.

Table 4. Results of model (logistic regression)

	Accuracy	Precision	Recall	ROC AUC	F1 Score
Train Set	88.50%	85.82%	83.85%	87.80%	84.83%
Test Set	88.35%	85.32%	83.75%	87.54%	84.53%

The model's predictions of the respective datasets denote the correctness of the accuracy. That shows the model can predict the class labels correctly for approximately 88.35% of the test set and 88.50% of the training set samples.

These values demonstrate that the model performs well in making accurate predictions, with a high percentage of correctly classified instances on both datasets.

In contrast, the "After PCA" section of the figure showcases the transformed data representation after applying PCA. The PCA is a potent dimensionality reduction technique designed to decrease the number of features while retaining as much of the data's variance as feasible. By projecting the original data into a space with a lower dimension, PCA identifies a new set of orthogonal axes as principal components. These components are arranged according to the variance they elucidate in the data.

Precision quantifies the ratio of true positive predictions among all positive predictions. In this case, the model has a precision of around 90.16% on the test set and 90.11% on the train set. This suggests that when the model classifies a sample as positive, it is accurate approximately 90% of the time. The high precision scores suggest that the model has a low false positive rate and effectively minimizes incorrect positive classifications.

Recall, also called sensitivity or true positive rate, represents the ratio of true positive predictions among all actual positive samples. The model's recall is approximately 91.15% on the test set and 91.40% on the train set, indicating that it can correctly identify around 91% of the positive samples. High recall values imply that the model has a low false negative rate and can effectively capture positive instances.

ROC AUC is a metric that evaluates the model's capacity to discriminate between positive and negative samples. The model achieves a ROC AUC score of about 0.8746 on the test set and 0.8762 on the train set, which indicates reasonably good discrimination power.

A ROC AUC score close to 1.0 suggests strong discriminatory capabilities, and the model's scores indicate its ability to effectively differentiate between positive and negative occurrences.

The F1 Score is the harmonic mean of precision and recall. It offers a balanced measure that takes into account both false positives and false negatives. The F1 Score is around 0.9065 on the test set and 0.9075 on the train set, showing that the model achieves a balanced measure between precision and recall. This balanced metric is especially beneficial when there is an imbalanced class distribution.

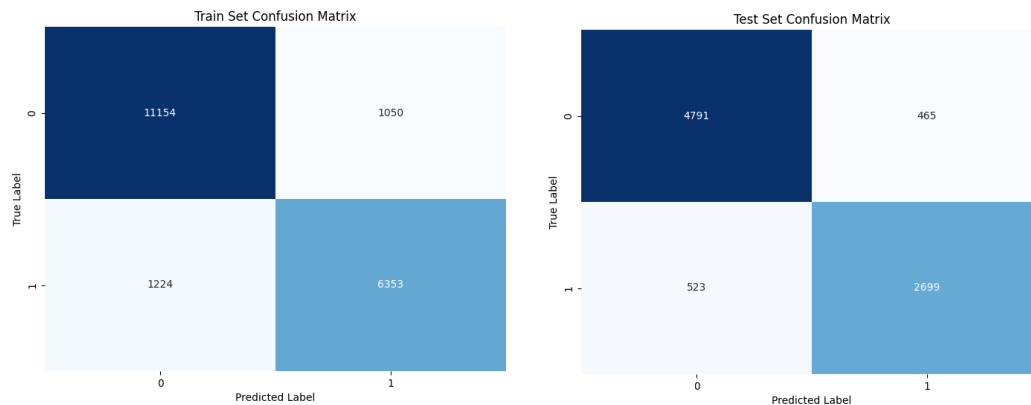


Figure 4. Logistic regression algorithm confusion matrix

In Figure 4, the model's performance on both the train set and the test set is presented in a clear and concise format. Figure 4 illustrates each set's counts of true positive, true negative, false positive, and false negative predictions. These values offer a detailed breakdown of the model's classification results on both datasets, providing insights into its ability to accurately identify instances from the positive and negative classes.

In the binary classification, evaluating a model's performance is of paramount importance in assessing its effectiveness. One of the key tools for evaluating classification models is the confusion matrix. In this article, we delve into an in-depth analysis of the performance of a logistic regression model using confusion matrices. We aim to understand how well the model can make accurate predictions and identify areas where it may be misclassifying instances.

The model of the test set exhibited a commendable performance, correctly predicting 91.15% of positive instances (TP) and 91.92% of negative instances (TN). However, it misclassified 465 samples from the negative class as positive (FP) and 523 samples from the positive class as negative (FN). The overall accuracy on the test set stood at an impressive 88.35%.

The results of the training set demonstrated similar trends to the test set, with the model achieving high accuracy (88.50%) and capturing 91.40% of positive instances (TP) and 91.37% of negative instances (TN) accurately.

The logistic regression model demonstrates commendable performance on both the test set and the train set, as indicated by the high accuracy, precision, recall, ROC AUC, and F1 Score values. These metrics offer a thorough assessment of the model's strengths and weaknesses in classification tasks and valuable insights for further model refinement and decision-making in real-world applications.

However, it demonstrated overfitting tendencies, misclassifying 1050 negative samples as positive (FP) and 1224 positive samples as negative (FN).

According to the results of confusion matrices, the logistic regression model demonstrated commendable performance on the test set, accurately predicting 91.15% of positive instances and 91.92% of negative instance and on the training set, the model achieved high accuracy (88.50%) and captured a significant portion of positive and negative instances correctly.

Overall, the model's ability to achieve high accuracy for both positive and negative instances on the test set suggests that it is performing well in distinguishing between the two classes.

The hypotheses were subjected to statistical testing using the Chi-square and Pearson correlation methods. The corresponding results for Hypothesis 1 and Hypothesis 2 are presented in Table 5.

The p-value reported as 0.0 in the Hypothesis 1 indicates that the observed result is highly statistically significant. In statistical hypothesis testing, a p-value of 0.0 means that the probability of obtaining the observed results, or even more extreme results, under the assumption of the null hypothesis (i.e., no association between the variables) is essentially negligible. Therefore, we reject the null hypothesis in favor of the alternative hypothesis, indicating that there is no significant relationship between gender and academic performance.

The correlation coefficient of 0.0847 between the variables indicates a very weak positive correlation. Although the correlation is statistically significant due to the large sample size, the magnitude of the correlation suggests that the association between gender and academic performance is minimal.

Table 5. Hypotheses testing results

	χ^2	df	p-value (Significance)	Correlation	Hypotheses
Hypothesis 1	1	2	0.0	0.0847	Reject H_0
Hypothesis 2	0.868	2	0.281	0.99	Accept H_0

The p-value reported as 0.281 in Hypothesis 2 indicates that the observed results are not statistically significant. As a result, we accept the null hypothesis instead of the alternative hypothesis. This implies a notable relationship between gender and exam characteristics, per the analyzed data. The correlation coefficient of 0.99 between the variables indicates a very strong positive correlation. It is important to note that while the correlation is statistically significant due to the large sample size, the magnitude of the correlation suggests that the association between gender and exam characteristics is indeed substantial. Female students are passing the exams with colloquia.

6. Limitations of the Study

The findings of this study are contingent on the quality and accuracy of the data used. The accuracy and completeness of the data gathered from the UIS are pivotal factors influencing the validity of the results. While the study aims to explore the relationship between the UIS and academic performance, there may be other confounding variables not accounted for in the analysis. Variables such as students' socio-economic backgrounds, prior academic performance, or external influences might influence the observed results. Due to the nature of observational studies, the analysis may identify correlations between variables, but it cannot establish causation definitively. While the study may suggest associations between the UIS and academic performance, it cannot conclusively determine the direction of causality. Logistic regression, PCA, and performance metrics have underlying assumptions that need to be met for accurate results. Failure to meet these assumptions, such as linearity, independence of observations, and normality, could impact the validity of the analysis.

The impact of the UIS on academic performance may vary based on institutional factors, such as teaching methodologies, support services, or extracurricular opportunities. The study's generalizability might be limited to specific educational settings or institutions with similar systems and resources. The study's scope is limited to the University Information System's impact on academic performance. The findings may not be directly applicable to other educational systems or settings outside the scope of this study.

Despite these constraints, the study provides valuable insights into the potential association between the UIS and academic performance.

7. Conclusion and Future Works

In this paper, we explored several data analysis techniques and performance metrics to evaluate the effectiveness of a logistic regression model. We utilized PCA to reduce the dimensionality of high-dimensional data, enabling better visualization and analysis of complex interdependencies between variables. The results after the PCA transformation demonstrated its capability to preserve a substantial portion of the data's variance.

The logistic regression model exhibited commendable performance on both the test and train sets, as evident by high accuracy, precision, recall, ROC AUC, and F1 Score values. These metrics collectively provided a comprehensive assessment of the model's strengths and weaknesses in classification tasks. Furthermore, the confusion matrices revealed the model's ability to accurately identify instances from the positive and negative classes.

Regarding our hypotheses, we found that there was no significant relationship between gender and academic performance, as indicated by the highly statistically significant p-value of 0.0 and a very weak positive correlation coefficient of 0.0847. Nonetheless, we noted a robust positive correlation of 0.99 between gender and exam characteristics, although this correlation did not achieve statistical significance, with a p-value of 0.281.

Overall, the research offers valuable insights into data analysis techniques, model evaluation, and the relationships between various variables. The findings contribute to the understanding of model performance and can aid in decision-making for real-world applications. It is essential to consider the limitations and further investigate the relationships identified to enhance the practical implications of the study.

The paper presents a comprehensive analysis, enriching our understanding of data analysis methods, logistic regression modeling, and the influence of gender on academic performance and exam characteristics. Future studies may build upon these findings and explore additional factors that could impact academic outcomes to provide a more holistic understanding of student performance.

References:

- [1]. J. Smith & A. Johnson, (2021). *Introduction to Systems Theory: Understanding Complex Systems*. New York: Academic Press.
- [2]. S. Thurner, P. Klimek & R. Hanel, (2018). *Introduction to the theory of complex systems*. Oxford University Press.
- [3]. L. Brown and M. Williams, (2022). The Role of Information Systems in Modern Universities. *Journal of Educational Technology*, 36(2), 145-160.
- [4]. Laudon, K. C., & Laudon, J. P. (2004). *Management information systems: Managing the digital firm*. Pearson Educación.
- [5]. Turban, E., Pollard, C., & Wood, G. (2018). *Information technology for management: On-demand strategies for performance, growth and sustainability*. John Wiley & Sons.
- [6]. I. Ali, A. Selim and E. Albayrak, (2023). *University Information System (UIS) User Manual*. International Vision University, Gostivar.
- [7]. J. A. O'Brien and G. M. Marakas, (2017). *Management Information Systems*. New York: McGraw-Hill Education.
- [8]. Stair, R. M. & Reynolds, G. (2020). *Principles of Information Systems (MindTap Course List)*. MA: Cengage Learning.
- [9]. Dawson, S., Heathcote, L., & Poole, G. (2010). Harnessing ICT potential: The adoption and analysis of ICT systems for enhancing the student learning experience. *International Journal of Educational Management*, 24(2), 116-128.
- [10]. Safiudin, A., Sulisty, M. E., Pramono, S., & Ramelan, A. (2020). The development of web-based Outcome Based Education information system. *Journal of Electrical, Electronic, Information, and Communication Technology*, 2(2), 61-64.
- [11]. Lee, M., Kim, H., & Kim, M. (2014). The effects of Socratic questioning on critical thinking in web-based collaborative learning. *Education as Change*, 18(2), 285-302.
- [12]. S. Palmer, (2013). Modelling Engineering Student Academic Performance Using Academic Analytics, *International Journal of Engineering Education*, 29, 132-138.
- [13]. Beranek, M., Bory, P., & Vacek, V. (2016). Platform for supporting student learning at unicorn college. *International Journal of Education and Learning Systems*, 1.
- [14]. Holderied, A. C. (2011). Instructional design for the active: Employing interactive technologies and active learning exercises to enhance information literacy. *Journal of Information Literacy*, 5(1).
- [15]. Abdous, M. H., Wu, H., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Journal of Educational Technology & Society*, 15(3), 77.
- [16]. Raji, M., Duggan, J., DeCotes, B., Huang, J., & Vander Zanden, B. (2017, October). Visual progression analysis of student records data. In *2017 IEEE Visualization in Data Science (VDS)*, 31-38. IEEE.
- [17]. Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems*, 161, 134-146.
- [18]. Mead, B. J., Estaphan, S., & Corrigan, G. (2023). Medical Students' Participation in Social Studying and Learning during COVID-19. *Education Sciences*, 13(4), 380.
- [19]. Chytrý, V., Řičan, J., & Medová, J. (2019). How teacher's progressiveness in using digital technologies influences levels of pupils' metacognitive knowledge in mathematics. *Mathematics*, 7(12), 1245.
- [20]. Harrell, I. L., & Bower, B. L. (2011). Student characteristics that predict persistence in community college online courses. *American Journal of Distance Education*, 25(3), 178-191.
- [21]. Houser, L. C. S., & An, S. (2015). Factors affecting minority students' college readiness in mathematics. *Urban Education*, 50(8), 938-960.
- [22]. Hu, S., McCormick, A. C., & Gonyea, R. M. (2012). Examining the relationship between student learning and persistence. *Innovative Higher Education*, 37, 387-395.
- [23]. Wu, S. J., Chang, D. F., & Sun, F. R. (2020). Exploring college student's perspectives on global mobility during the COVID-19 pandemic recovery. *Education Sciences*, 10(9), 218.
- [24]. Jelfs, A., & Richardson, J. T. (2013). The use of digital technologies across the adult life span in distance education. *British Journal of Educational Technology*, 44(2), 338-351.
- [25]. Liu, L. (2011). Factors Influencing Students' Preference to Online Learning: Development of an Initial Propensity Model. *International Journal of Technology in Teaching & Learning*, 7(2).
- [26]. Schumacher, P., Olinsky, A., Quinn, J., & Smith, R. (2010). A comparison of logistic regression, neural networks, and classification trees predicting success of actuarial students. *Journal of Education for Business*, 85(5), 258-263.
- [27]. Williams van Rooij, S. (2012). Open-source learning management systems: a predictive model for higher education. *Journal of Computer Assisted Learning*, 28(2), 114-125.
- [28]. Wilkins, S., & Balakrishnan, M. S. (2013). Assessing student satisfaction in transnational higher education. *International Journal of Educational Management*, 27(2), 143-156.
- [29]. Selimi, A., Saracevic, M., & Useini, A. (2020). Impact of using digital tools in high school mathematics: A case study in North Macedonia. *Universal Journal of Educational Research*, 8(8), 3615-3624.

- [30]. Selimi, A., & Üseini, A. (2019, April). Yenilikçi eğitim ile dijital yetkinlik ve girişimcilik becerilerinin geliştirilmesi—Kuzey Makedonya örneği. In *ICEB'19-International Congress of Economics and Business*, 11-13.
- [31]. Wu, J. P., Lin, M. S., & Tsai, C. L. (2023). A Predictive Model That Aligns Admission Offers with Student Enrollment Probability. *Education Sciences*, 13(5), 440.
- [32]. Laurell, J., Gholami, K., Tirri, K., & Hakkarainen, K. (2022). How mindsets, academic performance, and gender predict finnish students' educational aspirations. *Education Sciences*, 12(11), 809.
- [33]. Berei, E. B., & Pusztai, G. (2022). Learning through digital devices—Academic risks and responsibilities. *Education Sciences*, 12(7), 480.
- [34]. Fobel, L., & Kolley, N. (2021). Regional patterns of access and participation in non-formal cultural education in Germany. *Education Sciences*, 12(1), 13.
- [35]. I. T. Jolliffe, (2010). *Principal Component Analysis* (2nd ed.), New York: NY: Springer.
- [36]. Begg, R., Lai, D. T., & Palaniswami, M. (2007). *Computational intelligence in biomedical engineering*. CRC Press.
- [37]. Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd.
- [38]. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- [39]. Sarkar, D., Bali, R., & Sharma, T. (2018). *Practical Machine Learning with Python. A Problem-Solvers Guide To Building Real-World Intelligent Systems*, Apress Berkeley.