National College of Ireland

Evaluation of Sustainable Development Business Strategies of Higher Education Institutions using Data Mining and Machine Learning Techniques.

MSc Research Project
MSc in Data Analytics

Mary Ann Antony Kizhakechethipuza
Student ID: 23360895

School of Computing
National College of Ireland

Supervisor: Furqan Rustam

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mary Ann Antony Kizhakechethipuza
Student ID:	23360895
Programme:	MSc in Data Analytics **Year:** 2024-2025(September start)
Module:	MSc Research Project
Supervisor:	Furqan Rustam
Submission Due Date:	August 11th, 2025
Project Title:	Evaluation of Sustainable Development Business Strategies of Higher Education Institutions using Data Mining and Machine Learning Techniques.
Word Count:	6842 **Page Count:** 25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Mary Ann Antony Kizhakechethipuza

Date: August 11th, 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple	□

copies).	
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Evaluation of Sustainable Development Business Strategies of Higher Education Institutions using Data Mining and Machine Learning Techniques

Mary Ann Antony Kizhakechethipuza
23360895

Abstract

Sustainable Development Goals when introduced by United Nations provide a much-needed structure to streamline sustainability-based efforts. Higher Education Institutions (HEIs) provide a conducive environment where research related to sustainability can flourish. They are in a unique position to lead by example and build knowledge on sustainability that can be harnessed by future generations. This project pursues the direction of using THE-IR data to evaluate the sustainable development-based strategies of HEI using well established machine learning models like KMeans Clustering, Logistic Regression and Random Forest Classifier. The feasibility of pursuing the addition of a technological dimension of sustainability is also addressed using Bibliometric analysis. Key findings include that THE-IR data can be used in evaluation of sustainability-based strategies as it works well with traditional models employed in the education data mining domain. Digital Transformation emerged as a major theme surrounding technological sustainability. More research is needed to identify if there is feasibility in including the technological dimension to the traditional triple bottom line framework.

1 Introduction

The introduction of Sustainable Development Goals (SDGs) by United Nations brought a renewed structure to businesses interested in focussing their efforts on sustainability. The SDGs immediately gained momentum because of their correlation with the three dimensions of sustainability namely social, economic and environmental dimensions. Basheer et al. (2025) outlines that Higher Education Institutions (HEIs) are in a unique position to advance specific SDGs like SDG 4 (Quality Education), SDG 11 (Sustainable Cities and Communities), and SDG 13 (Climate Action) by helping build the knowledge necessary for application in real world. Avelar and Pajuelo-Moreno (2024) explores the three streams through which HEIs can promote the advancement of SDGs namely Research, Outreach and Teaching. The research particularly highlights the significant impact HEIs have in shaping future generations with professional skills necessary for addressing SDGs in the work they do, potentially shaping the future of economy, society and the environment. Hence, now more than before, it is necessary to understand the efforts and strategies employed by HEIs in advancing sustainability in their institutions.

Deda et al. (2025) discuss the sustainability assessment and subsequent analysis that was

conducted using 2022 Times Higher Education Impact Ranking (THE-IR) data on top 25 HEIs. The research highlights that while THE-IR's transparent methodology and its approach towards scoring the impact of HEIs towards all the SDGs are comprehensive for sustainability-based research objectives, there are limitations in the data itself like the varied representation across regions, scoring based on self-reported data and the methodology's focus on research intensive institutions. Ordonez-Ponce et al. (2024) also shows similar approach of using THE-IR data from 2019 to 2023 to assess sustainability performance in SDGs of 27 Canadian HEIs. The overall results from the analysis showed that Canadian HEIs were performing well on the social and economic SDGs while a constant decline in the environmental SDG (SDG 13) was witnessed. While both this research focussed on specific SDGs as representatives for each of the sustainability dimensions, this research accounts for the interlinkage between the SDGs in progressing the three dimensions using THE-IR data.

Finally, technology has become an integral part of HEIs business and management. The pandemic initiated a lasting change in the traditional channels of education delivery cementing the role of Technology as an important tool to achieve long-term sustainability. Nunez-Naranjo et al. (2025) discusses the real-world implications of technology inclusion in education in Latin America. The research highlights that the pandemic accelerated technological education learning rate, increasing the dependency on using technology for achieving continuation of business. Hence it is important to gain visibility on both the efforts employed to achieve progress in the SDGs while understanding the current research that exists on the technological dimension of sustainability such that a convergence can be reached for the inclusion of this dimension with the existing three pillars of sustainability. Tafese and Kopp (2025) have contributed significantly in this area by executing a bibliometric analysis to understand recurring research themes in different subject areas like Engineering, Energy, Environmental Science etc for sustainable development. Hence the research question pursued for this project includes:

Research Question 1 (RQ1): What is the impact of using bibliometric analysis to understand the major themes in analysis of technology as a possible dimension of sustainability?

Research Question 2 (RQ2): What is the impact of using THE-IR data to predict the sustainable outcomes of HEI's business strategy?

This research aims to further the knowledge in the analysis of HEI's efforts towards sustainability by using established SDG framework and publicly available data like THE-IR. In addition, the novelty of this research is the analysis done on the possibility of achieving a convergence in the inclusion of the technological dimension to the existing three pillars of sustainability is also addressed through bibliometric analysis of existing research related to the subject. This document is structured to include the following sections

- **Section 2- Related Work:** The literatures that were used to decide the direction of the research project are discussed here and are included as subsections where applicable.

- **Section 3- Research Methodology:** This section outlines the methodology used along with the different steps involved with data collection, cleaning and preprocessing the data for both research questions. The selection of algorithms is also briefly discussed here.
- **Section 4- Design and Implementation**: This section outlines the design of the analysis, the algorithms used for analysis as well as the combination of techniques used to reach at the results.
- **Section 5- Evaluation**: This section outlines the interpretation of the results achieved from section 4 and the suitability of using the methodology and data towards further research
- **Section 6- Discussion:** This section summarizes the key findings learned through execution of the design and its implementation on the chosen dataset.
- **Section 7- Conclusion and Future work:** This section summarizes the overall success of the research methodology in satisfying the initial objectives of the research and the research questions. It also addresses possible future directions for the research.

2 Related Work

Different literatures were reviewed to accommodate the different areas of this research project. Each of these areas are explained as a subsection in the content below and associated literatures relevant to these areas are discussed individually and summarised to show why a particular direction was chosen.

2.1 Sustainable Development Business Strategies

Traditionally, Sustainability has been defined to include three dimensions namely Social, Economic and Environmental dimensions. This is identified as the Triple Bottom Line framework and it provides a rounded approach to understand, channel and measure efforts towards sustainable development in a meaningful way. Nogueira, Gomes and Lopes (2025) research findings detail that aligning a company's strategic objectives towards the three dimensions leads to significant economic development, positive effect on the company's business performance and ability in achieving long-term sustainability. Kusmendar, Asih and Normasari (2025) studied in detail the implementation of TBL approach across different sectors. The study highlights the key gaps in literature in being able to measure the consistent implementation of TBL across industries. Both these literatures through bibliometric analysis of several other literatures highlight that TBL conceptual framework is well accepted and recognized across business as an grounded approach to understand and measure sustainability. The introduction of 17 Sustainable Development Goals(SDGs) by United Nations provided clear points to match sustainability based efforts to measuring success/progress in achieving them. The strategies that contribute to progressing SDGs and their implementation varies across industries, but for Higher Education Institutions(HEIs), one possible route to measure individual and global progress in these SDGs are through

sustainability rankings. Calderon (2023) examines three international sustainability rankings namely Universitas Indonesia GreenMetric World University Ranking (GreenMetric), Times Higher Education Impact Ranking (THE-IR), QS Sustainability Rankings (QS-SR), and the Sustainability Tracking Assessment & Rating System (STARS) schema. These rankings, each through their unique approach enable HEIs to shed visibility on their efforts in achieving sustainability to the world. Urbano et al. (2025) studied Times Higher Education Impact Ranking(THE-IR)in detail and analysed the extent to which these rankings capture the current sustainability level of an HEI. The literature also identifies that THE-IR provides a robust picture of the HEIs contribution towards individual SDGs but there can be disparities in the data based on HEIs participation and transparency in sharing information on their efforts. The rankings can be accessed publicly and the data capture is in consistent format(Ranking by SDG) every year making it easy to access and analyse the data for the purpose of this project. Hence THE-IR was chosen to pursue the analysis related to the project.

Higher Education Institutions(HEIs) are enablers of knowledge. Shawe et al. (2019) studied 16 HEIs (7 Irish and 9 International HEIs) to find that HEIs play an important role in enabling sustainable development. Though the actual implementation of sustainable development based strategies showed a mismatch with the number of initiatives outweighing established policies, HEIs provide the resources and environment necessary to build knowledge and research focused on tackling challenges related to sustainability.

2.2 Technological Dimension for Sustainability

Technology is a strong partner for any business, necessary to build capacity and resilience in the face of a changing environment. Okur et al. (2025) discuss how digital transformation has enabled HEIs to sustain and deliver education during major disruptions like the pandemic, which later changed the traditional way to gain education. The research also highlights that there is an imbalance in the weight placed by each of the three sustainability dimensions when deciding which technology to implement for specific purposes with economic dimension being more important in majority of the cases. Basheer et al. (2025) outlines how emerging technologies have enabled the need for their adoption both for their utility in streamlining operations and increasing efficiency but also to be able to sustain to new demands of the society fuelled by the boom in using these technologies. This research sheds light on the interconnectedness between the three dimensions of sustainability on technology adoption both within and outside the HEI. Hence, in the current changing world, the study of sustainability wouldn’t be complete without the addition of the technological dimension in addition to the existing three dimensions of sustainability.

2.3 Data Mining and Machine Learning application in Education Industry

Correa-Peralta, Vinueza-Martínez and Castillo-Heredia (2025) discuss the evolution of Business Intelligence and data analytics application in the management of HEIs. The research highlights techniques like logistic regression, clustering, decision trees, random forests and

artificial neural networks to be most applied in Educational data mining discipline which majorly focuses on improving student experience. In the domain of student performance/dropout prediction, Cruz and Lumauag (2024) also second the findings of the previous research findings where clustering, logistic regression and Random Forests are among the top approaches used in prediction of categorial variables. As these literatures have concluded their findings based on extensive literature review relevant to their study and domain, a similar approach has been chosen to analyse THE-IR dataset for the purpose of this project.

3 Research Methodology

The research methodology, tools employed, and data collected varied based on the research question pursued. KDD process was predominantly used and adapted where necessary to execute the different steps involved in the research.

3.1 RQ1- Bibliometric Analysis Methodology

The methodology for RQ1 was a mixed methods approach where all articles that matched the search criteria was retrieved from Scopus database. This marked the beginning of data collection for Bibliometric analysis. The metadata of the search results was exported as an excel file. Applying PRISMA technique on the data according to predefined criteria, this data was further refined to just the content selected for thematic analysis. The figure below explains the different steps followed as part of the RQ1 methodology with tools/software used in each step.

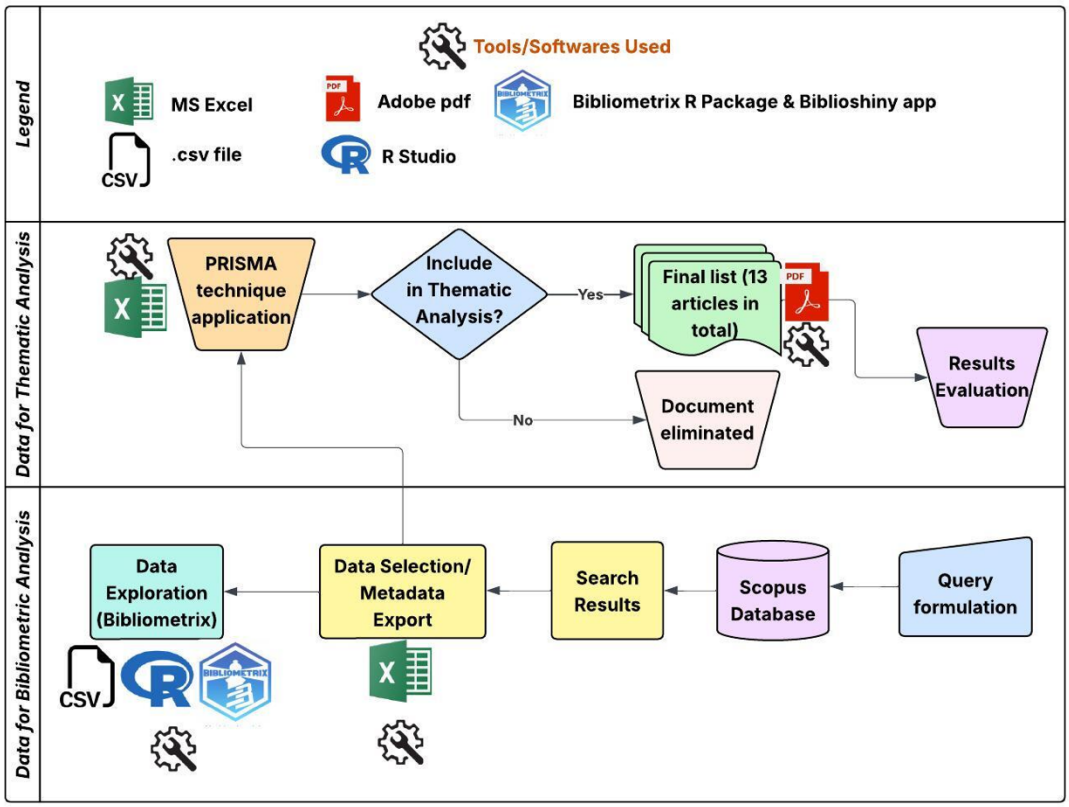


Figure 1: RQ1- Bibliometric Analysis and Thematic Analysis Methodology

3.1.1 RQ1- Bibliometric Analysis Methodology- Data Collection and Cleaning

Scopus database was used as the main source of data for the bibliometric analysis. The search criteria were iterated multiple times to reach the final query structure. The finalized search criteria are shown in the table below.

Table:1 Search criteria for Bibliometric Analysis

Search Within	**Search Documents**
Title, Abstract, Keywords	(technological AND sustainable AND development)
AND (operator)	
Language	(english)
AND (operator)	
Abstract	(sustainable AND development)
AND (operator)	
Title, Abstract, Keywords	(Higher education institutions)

The initial search was executed in May 2025. The elements of the search criteria were finalized through iterations and the finalized search results were extracted on June 4th, 2025, to an excel spreadsheet directly from Scopus database. The search results contained a total of 197 results with different information available like abstract, DOI, author Keywords, Publication date, Publisher etc. Trevisan et al. (2023) outlines the application of PRISMA technique to refine the list of literature selected towards content/thematic analysis after extraction. The steps in this research were adapted and used as a guide in the application of this technique for fine tuning the selection of literature towards thematic analysis. The table below shows the different elements/filters used as part of PRISMA technique to finalize the literature for content analysis. The final list included 13 literatures and all of them were open source, available in pdf format for further analysis using the principles of thematic analysis in Adobe pdf software. This direction was pursued due to the list being small and manageable without requiring professional software like Atlas to generate codes for mapping the majorly occurring themes in the literature.

Table 2: PRISMA Technique applied to shortlist literature for content analysis

Filter	Reasoning	Result	Stage
Source title- Sustainability(Switzerland)	This source research is theoretical with little application of Data Mining and Machine Learning techniques in their analysis. Hence all of the literature from this source(25 in total) was eliminated from content analysis.	List reduced to 172 literatures	**Screening**
Year - 2025, 2024, 2023 only	The initial list had all literatures starting from 1998 till 2025. As the SDGs were introduced only in 2015 with the aim to include it as part of the 2030 agenda for sustainable development, literature from 2023 onwards was selected.	List reduced to 102	

Document Type- Article, Review only	All literature that was an article or review was only selected from the different types available namely Book, book chapter, Conference paper, conference review, Erratum	List reduced to 80	
Author Keywords- Non blanks only	For 7 literatures, author provided keywords were missing, hence was eliminated from selection	List reduced to 73	
Combination of words in title and abstract	Digital/Digital transformation, Sustainability/Sustainable development, AI/Machine Learning/Deep Learning, Education, SDG were the different words that were searched manually in the abstract and title to shortlist the final list for content analysis. All of the literature were from 2025 and the list size reduced significantly.	List reduced to 13	**Final**

3.1.2 RQ1- Bibliometric Analysis Methodology- Tools used

There are different tools and applications available to perform bibliometric analysis. Bibliometrix library in R studio was used to execute the analysis for this project. The other tools like VOS viewer were pursued for the same purpose but had compatibility issues with the computer's operating system. All analysis was performed using Biblioshiny app which can be initiated from R console. The app opens as a new tab in the browser with IP address http://127.0.0.1:5032. A stable version of RStudio is a prerequisite to be able to install the Bibliometrix library. The installation and initiation procedure to follow is clearly outlined in bibliometrix.org. All analysis was carried on a mac operating system-Sequoia 15.5. The figures below show the initiation of Bibliometrix and Biblioshiny from RStudio console window.

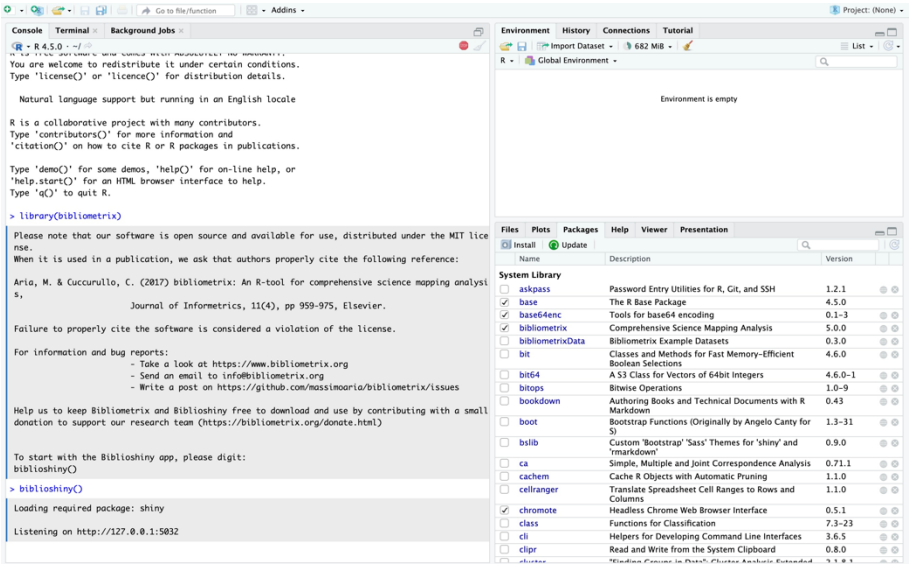


Figure 2: Bibliometrix Library and Biblioshiny app initiation from RStudio Console

3.2 RQ2- THE-IR Analysis Methodology (2019 and 2025 dataset):

For RQ2 methodology, quantitative approach was followed. The data collection process was initiated by manual extraction of THE-IR data from the Times Higher Education website. The rest of the steps followed mirrored the KDD process like shown in the figure below. The same methodology was followed for both 2019 and 2025 datasets. The rankings for individual SDGs were provided separately on the website. Hence, for each year, each of the 17 SDGs had to be extracted separately and then manually merged to form the master dataset on which additional features were added. Additionally, the number of HEIs participating per SDG varied for the same year. Hence the merging of each SDG separately to the master data sheet required matching of the names of the HEIs in both list with the excess HEIs added to the master list directly as a new entry with their corresponding scores.

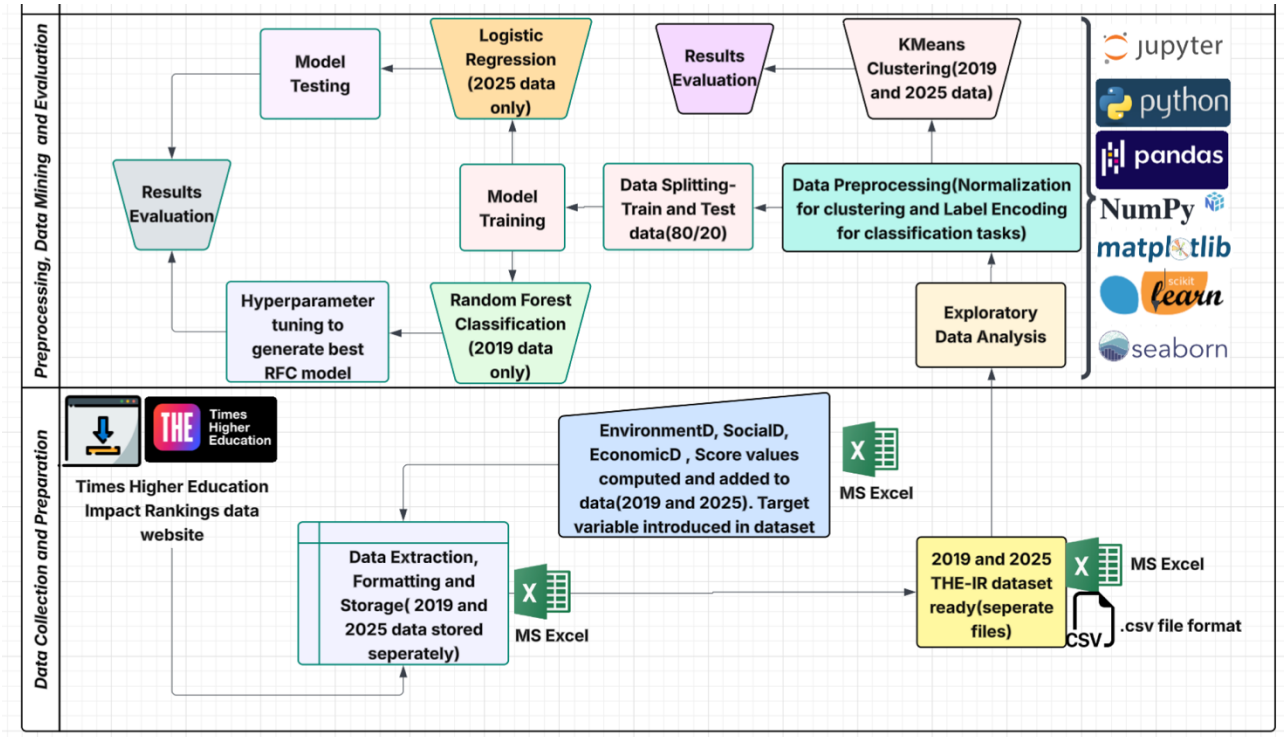


Figure 3: RQ2- THE-IR data analysis Methodology (2019 and 2025 data)

3.2.1 RQ2- THE-IR Analysis Methodology- Data Collection, Cleaning and Feature addition

THE-IR data is available publicly to download from the website-timeshighereducation.com under the webpage Impact Rankings. Data is available from the year 2019 up to the year 2025. The rankings for all the SDGs are included in a tabular format. The data collection was the most time-consuming step among the other steps involved with the project. In the interest of time, the data for year 2019 and 2025 was manually extracted to an excel spreadsheet. The data was then cleaned of any formatting, links attached to names of the universities. Additional columns were added like "Region" where the universities were mapped to 6 regions namely Africa, Asia, Europe, Middle East, North America, South America and Oceania. The breakdown of the regions with the countries is shown in the table below.

Table 3: Breakdown of the THE-IR dataset by region and year

Year	Region	Count
2025	Africa	287
2025	Asia	1069
2025	Europe	495
2025	Middle East	361
2025	North America	113
2025	South America	165
2025	Oceania	37(Total=2527 rows)
2019	Africa	32
2019	Asia	185
2019	Europe	152
2019	Middle East	58
2019	North America	60
2019	South America	47
2019	Oceania	23(Total = 557 rows)

Regions Asia and Europe had maximum representation in both years. In addition, there were a few universities who had rankings only for a few SDGs. The missing fields were filled with null values (0) throughout the dataset and later confirmed in the implementation stage with exploratory data analysis. The header file had different format of labels for the SDGs. There were all edited to follow a single format for the labels- Single word labels with no special characters. The cleaned 2019 dataset had 557 unique instances(rows) with 22 columns whereas as the cleaned 2025 dataset had 2927 unique instances (rows with same list of 22 columns. For both year data, the three dimensions of sustainability were included by creating three additional columns using the methodology addressed in Stockholm Resilience Centre webpage

Figure 4: SDG's wedding cake methodology

This methodology views the SDGs as interconnected with progress in specific SDGs contributes to specific dimensions of sustainability. Applying this methodology to both years of data,

- "EnvironmentD" column representing to the Environmental dimension was calculated as the average of the impact values recorded for SDG 6, 13, 14 and 15.
- Similarly, "SocialD" representing the social dimension was calculated as average of the impact values recorded for SDG 1, 2, 3,4,5,7,11,16.
- For the Economic Dimension, column "EconomicD" was calculated as an average of the impact values recorded under SDG 8,9,10,12.
- Finally, another column "Score" was created which was the average of the values recorded under each of these columns and SDG 17. All the averages were rounded to a single decimal point to keep the data consistent across the columns. This value was chosen as the Sustainability score for that HEI.
- This method helped to resolve two issues:
  - For HEIs that has sparse representation for most of the SDGs were able to be represented with a defined value and not null value for the analysis.
  - This helped to train the model based on a unique score per HEI and to set a condition for classification based on this feature as there wasn't a feature available previously that could be depended on to train the model effectively for prediction purposes.

SMOTE technique was not employed in this case to balance the representation of the regions as this data was compiled based on information provided by actual HEIs focussing on sustainability. Any addition of synthetic data for the purpose of balancing the dataset would be misrepresentation of reality.

Finally, another column named- "Target" was added onto the data with 2 categories namely 0 and 1. For 2025 data,

- category 0 included all values of score <=32 and
- category 1 included all values of score >32 up to the maximum value of 90.3.

Similarly for 2019 data,

- category 0 included all values of score <=40 and
- category 1 included all remaining values up to the maximum value of 70.7.

This is the target variable that is being used for classification and prediction purposes. Category 0 denoted "not sustainable HEIs" and Category 1 denoted "sustainable HEIs" for both years of data.

3.2.2 RQ2- THE-IR Analysis Methodology - Tools used

Data extraction and cleaning was solely done using Microsoft Excel. The values for the three dimensions were calculated by applying function to one cell of the row and copying it across the entire column. Excel was preferred as it was easier to clean and faster to process both the data than using python. The cleaned data was then converted to csv format with UTF-8 encoding to be able to use in jupyter notebook for further analysis. The analysis was performed using python programming language. Python libraries like pandas, NumPy were used to export the data into the dataframe and libraries like sklearn and seaborn was used to choose the model, fit the model to the data, train the model and generate metrics that was later interpreted during results evaluation.

4 Design Specification and Implementation

The design and implementation of the research used for each RQ has been discussed separately in the subsections. They are clubbed together to make it easy to follow through with the different steps associated with each RQ.

4.1 RQ1- Bibliometric Analysis

From the Scopus database, the finalized search results were downloaded as a csv file. This csv file was uploaded in biblioshiny app initiated from RStudio. The default settings were used for this purpose. The csv data was uploaded to biblioshiny app using the Import or Load option. The picture below shows the input setting before initiating the analysis using "Start".

Figure 5: Biblioshiny Data upload using Import or Load function

As a next step, the system provided a report on the quality of the different parameters included in the original dataset. I proceeded with the analysis using the exact same quality of the parameters as it was directly exported from Scopus.

Completeness of metadata -- **201** docs from **Scopus**

Metadata	Description	Missing Counts	Missing %	Status
AB	Abstract	0	0.00	Excellent
DT	Document Type	0	0.00	Excellent
SO	Journal	0	0.00	Excellent
LA	Language	0	0.00	Excellent
PY	Publication Year	0	0.00	Excellent
TI	Title	0	0.00	Excellent
TC	Total Citation	0	0.00	Excellent
AU	Author	4	1.99	Good
C1	Affiliation	6	2.99	Good
CR	Cited References	12	5.97	Good
DI	DOI	26	12.94	Acceptable
DE	Keywords	41	20.40	Poor
RP	Corresponding Author	66	32.84	Poor
ID	Keywords Plus	110	54.73	Critical
WC	Science Categories	201	100.00	Completely missing

Advice Report Save Close

Figure 6: Biblioshiny system report on data quality of the dataset

As a next step, in the Settings bar, I increased the figure size and resolution to maximum values provided such that the figures exported have higher resolution. After this each of the reports generated under different sections were added to the report using the “add” button and then the report exported from the “Report” tab. The figures were downloaded as n excel spreadsheet which had both the data and figures in separate tabs.

The list below summarizes the tools used in this design and implementation:

- **Microsoft excel**- Data collection, cleaning and data processing to different formats like csv to enable upload of dataset to other applications.
- **Scopus database**-Main source of data for this analysis
- **RStudio**- This software was already downloaded to my computer. Bibliometrix library was installed and Biblioshiny app was initiated from this tool.
- **Bibliometrix library and Biblioshiny app**: Bibliometrix library was first initiated followed by biblioshiny app which opened as anew tab in the browser. Default settings was used in both cases.

4.2 RQ1- Thematic Analysis for Technological dimension addition

The finalized list of literatures was individually downloaded in pdf format and manually read to understand the different themes that were discussed under the technological sustainability. No additional software was used for this purpose except Adobe pdf viewer due to the size of the literatures that we included and also to simplify the thematic analysis process.

4.3 RQ2- THE-IR data analysis using machine learning applications

As discussed in the literature review- section 2.3 and according to Cruz and Lumauag (2024), clustering, logistic regression and Random Forests are among the top approaches used in prediction of categorial variables. The THE-IR data was modelled in a similar fashion with feature ”target” being categorical. The intention was not to restrict the analysis to the 3 models selected, rather to use the best approach that would fit the structure of the data, supported by evidence in previous research work. In the THE-IR data, the values ranged between 0 and 100 for all SDGs. This structure resembled student performance scores datasets where there is a similar range setting. Hence Logistic regression and Random Forests was chosen to perform the analysis.

Additionally, null values were prominent in THE-IR data. Not all SGDs had scores available for the same set of HEIs. In other words, majority of the HEIs had scores only for a few SDGs and missing values for the rest making the presence of null values prominent in the dataset. Based on a quick calculation for the 2019 dataset, the presence of null values were between 10% and 58% of the total instances for all SDGs except for SDG 17. Similarly for the 2025 dataset, the presence of null values were between 10% and 66% of the total instances for all SDGs except for SDG 17. As Artificial Neural networks(ANNs) do not work

well with missing values and any imputation strategies used would alter the actual score received by the HEI in reality, analysis based on deep learning models were not pursued.

For both datasets, Clustering was done initially using KMeans Clustering technique. This route was decided based on the sckit-learn algorithm cheat sheet. As the number of categories were known and the size of both datasets were less than 10000 instances, KMeans was chosen as the clustering technique for both the datasets. The table below summarizes the parameters of the final model that was designed and used on the dataset.

Table 4: Machine Learning model design for THE-IR data analysis

Algorithm	Design Parameters of finalized model	Dataset
KMeans Clustering	**Programming language:** Python **Application:** Jupyter Notebook **Data preprocessing**: Normalization using MinMaxScaler **Train-test split:** 80/20 **Random_state:** 23360895 **k =n_clusters** = 6 **n_init**='auto'	2019 THE-IR
KMeans Clustering	**Programming language:** Python **Application:** Jupyter Notebook **Data preprocessing:** Normalization using MinMaxScaler **Train-test split:** 80/20 **Random_state:** 0 **k =n_clusters** = 3 **n_init**='auto'	2025 THE-IR
Random Forest Classifier	**Programming language:** Python **Application:** Jupyter Notebook **Data preprocessing:** Label encoding **Train-test split:** 80/20 **Random_state:** 23360895 **Features included for training set:** columns Name, country, region, Sdg 3,4,5,8,9,10,11,12,13,16,17, Target **Best hyperparameters:** {'max_depth': 17, 'n_estimators': 234} Default values were used for all other hyperparameters	2019 THE-IR
Logistic Regression	**Programming language:** Python **Application:** Jupyter Notebook **Data preprocessing:** Label encoding **Train-test split:** 80/20 **Random_state:** 16 **max_iter:** 1000 **Features included for training set:** columns	2025 THE-IR

	Name, country, region, EnvironmentD, SocialD, EconomicD, Target All other hyperparameters like penalty, dual, fit_intercept etc were used with default values.	

The main libraries used in both cases include pandas and NumPy to load data into the dataframe. Additionally, library seaborn, matplotlib and sklearn were used to visualize the results from implementing the models.

5 Evaluation

5.1 RQ1 Bibliometric Analysis and Thematic Analysis

There was a variety of results that were extracted from biblioshiny app. But only a few key findings are discussed here. One of the key findings were related to keywords. It was seen that the most used keywords were higher education, higher education institutions, sustainable development etc. But keywords like technological development, digitization were used less frequently in the literatures chosen. This was a drawback as most of the literatures didn't include the content necessary to gauge the different emerging themes related to technological dimension of sustainability. The finding of the Thematic analysis was that Digital Transformation was the major theme in all the literatures and emphasis was given on the adoption of new technologies like Blockchain in the educational sector. There wasn't enough information or previous work done to explore the inclusion of technological dimension but based on the content, technology was generally used to achieve sustainability and optimize operations without much emphasis placed on the assessing the three dimensions of sustainability for an adopted technology. Hence, it can be concluded that more research is required in this theme to be able to pursue this research question in future.

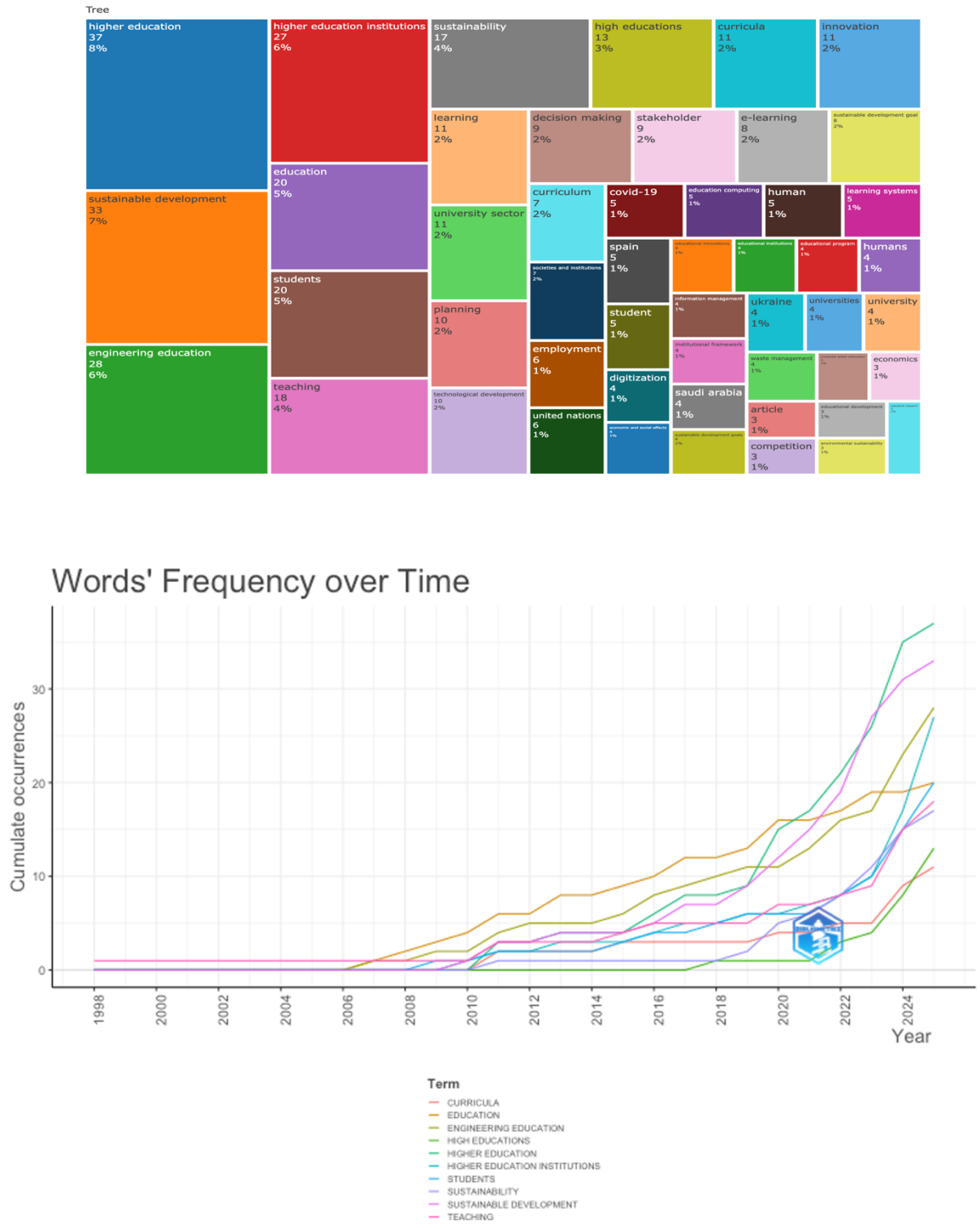


Figure 7: Word Frequency plot from Biblioshiny for the chosen dataset

5.2 RQ2 -KMeans Clustering/ 2019 and 2025 THE-IR data

The final model was selected based on the value of the silhouette_score. The lower the score the better the model's fit for the data. For 2019 data, the final model that worked best on the data was k=6 with a silhouette_score of 0.295. Figure 5 below shows the clusters division for Environment dimension, social dimension and Economic Dimension along with the 6

clusters. It is worth noting that 4 among the 6 clusters have similar distribution. Figure 6 shows the clusters division for Environment dimension, social dimension and Economic Dimension along with the 3 clusters. This model had a silhouette_score of 0.307. It is also worth noting that 2 of the cluster show similar distribution when compared to the third cluster. It can be inferred from both these plots that, the sustainability efforts reflected in the THE-IR data is that there is overlap in the sustainability efforts across the globe without a clear distinction in the clusters.

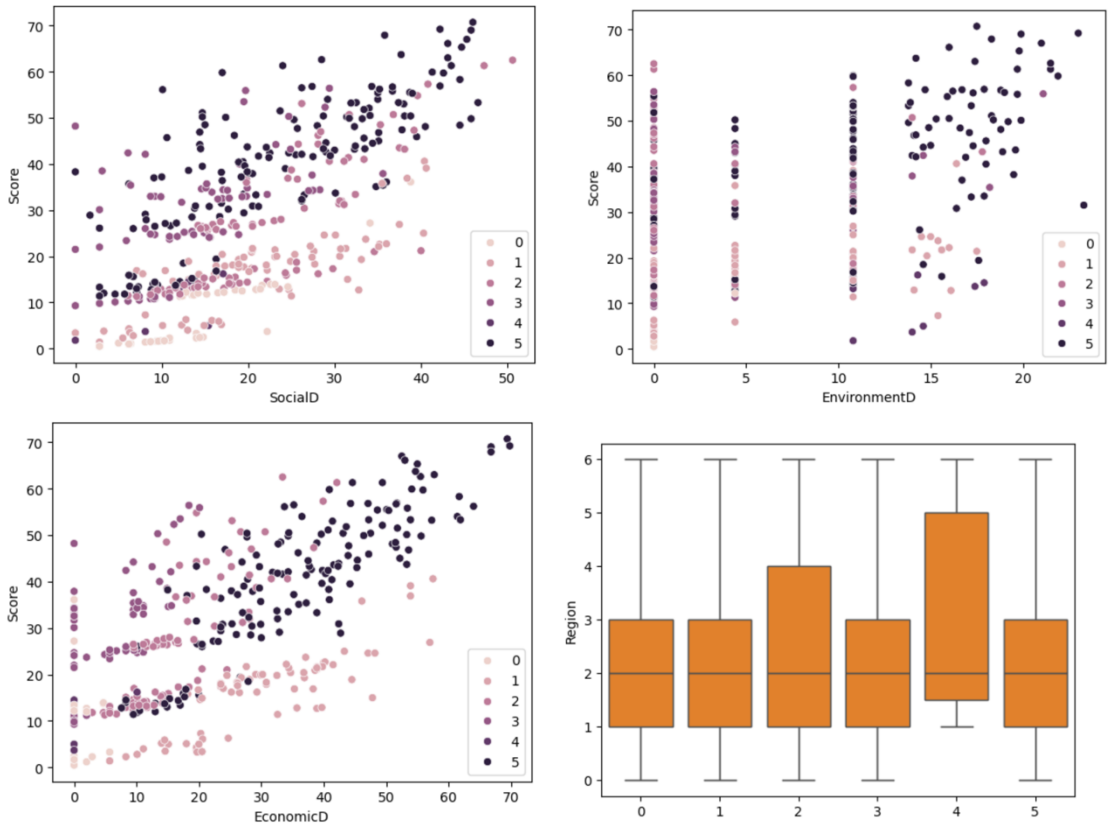
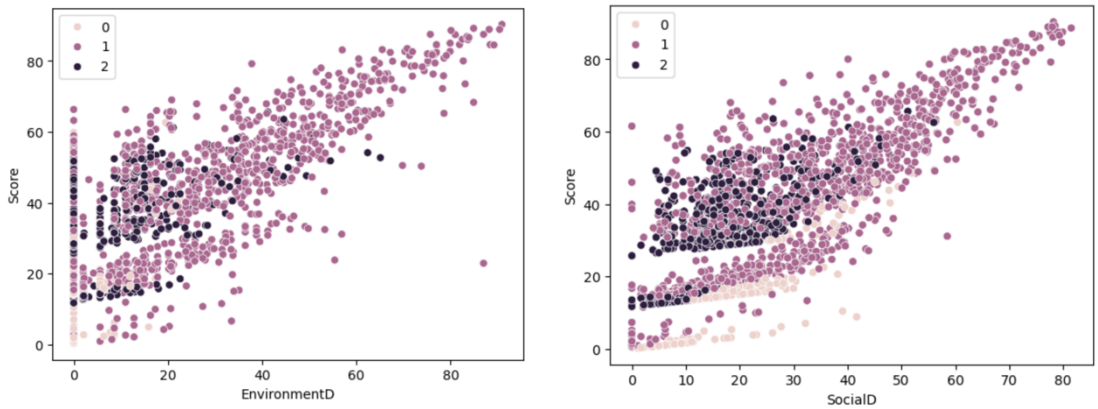


Figure 8: KMeans final model output for 2019 THE-IR data



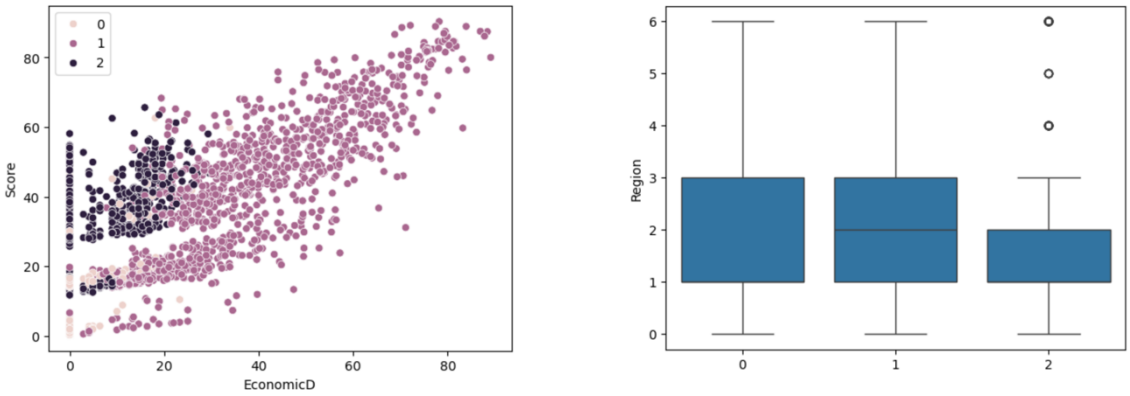
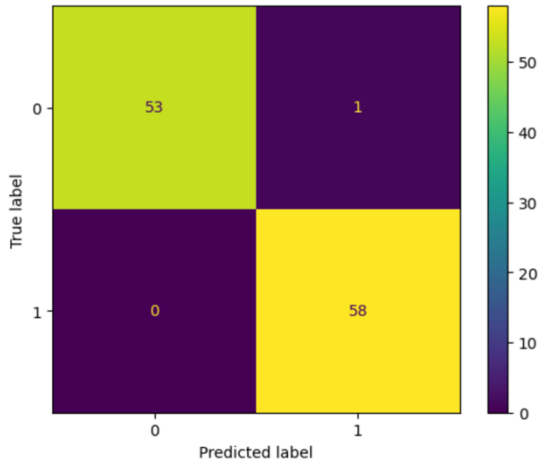


Figure 9: KMeans final model output for 2025 THE-IR data

5.3 RQ2 -Random Forest Classification/ 2019 THE-IR data

The random forest classifier showed high accuracy, precision and recall when applied to the features selected. These features were selected to ensure that the model had higher accuracy as there was no information biasing the performance of the model. The Feature "Score" though used to set the target category for the data was dropped from the training data to ensure the model only trained on the impact values of SDGs. The figure below shows the feature importance, confusion matrix and classification report of this model.



	precision	recall	f1-score	support
0	0.95	0.97	0.96	110
1	0.97	0.95	0.96	113
accuracy			0.96	223
macro avg	0.96	0.96	0.96	223
weighted avg	0.96	0.96	0.96	223

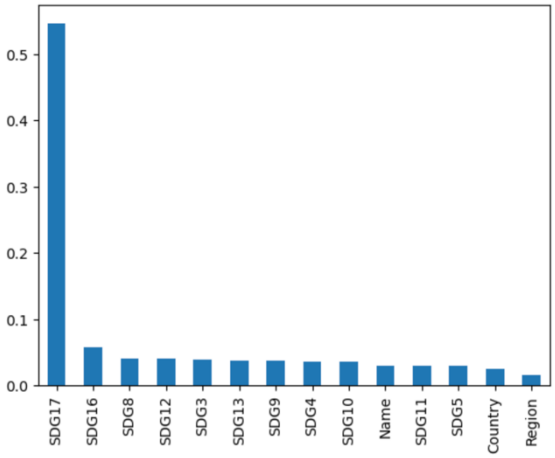


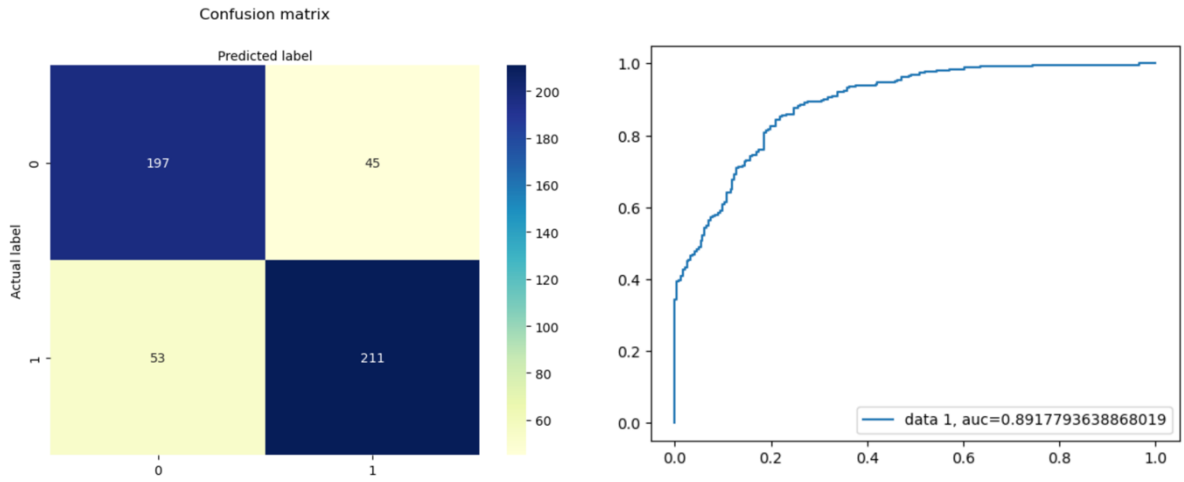
Figure 10: Random Forest Classifier output metrics for 2019 THE-IR data

Based on the metrics it can inferred that the model classified the data with an accuracy of 96%. The other metrics like precision, recall was also high at 95% and 97% respectively. It can be concluded that this model was a good fit in classifying the 2019 THE-IR data. Also, among the key features that the model utilized in classification, SDG17 was SDG16 were the top contributors. This is possible because based on the data, only these two columns had more representation from all the HEIs compared with the other impact values. This can viewed as sparse representation rather than bias in the data which is common in real-world data.

For Random Forest Classifier model(2019 THE-IR data), the best hyperparameters values was programmatically retrieved using rand_search and the best_rf model was created using these hyperparameters and then the confusion matrix, classification report generated to evaluate the performance of the model. A second random forest model was also built with test_size=0.40, stratify =y in the train-test split parameters and classification report was generated and compared with the best_rf model. The second model showed slight reduction in the accuracy, recall, precision and f1 score. Hence the best_rf model was finalized for this dataset.

5.4 RQ2 -Logistic Regression/ 2025 THE-IR data

Logistic regression model showed high accuracy of 81% when applied to 2025 THE-IR data. Other metrics like Precision and Recall also showed good levels at 79% and 81% respectively. The figure below shows the output from the model as applied to the 2025 THE-IR dataset. Here the features were selected were the three dimensions- EnvironmentD, SocialD and EconomicD along with the other features.



	precision	recall	f1-score	support
not sustainable	0.79	0.81	0.80	242
sustainable	0.82	0.80	0.81	264
accuracy			0.81	506
macro avg	0.81	0.81	0.81	506
weighted avg	0.81	0.81	0.81	506

Figure 11: Logistic Regression output metrics for 2025 THE-IR data

It is also worth noting that the model showed higher precision in predicting "sustainable" HEIs than higher recall in predicting "not sustainable" HEIs. But based on metrics this model is a good fit in predicting categorical variables and fits well for 2025 THE-IR dataset

For logistic regression model(2025 THE-IR data), a second model was constructed with train-test split parameters namely test_size = 0.30, random_state= 23360895 and shuffle= True. The classification report was generated and compared with the finalized model. There was a slight reduction in the accuracy, precision, recall and F scores for the second model. Hence the first model was chosen as the finalized model.

6. Discussion

For RQ1, based on the keywords evaluation results, it can be deduced that current research, though includes concepts like digitization is more focussed on the use of technology to improve efficiency in different management aspects of HEI. This could include employing technological development to improve operational efficiencies, reduce costs, improve sales through targeted marketing etc. There was lack of evidence in the literatures reviewed regarding conceptualizing technology and its application as a dimension of sustainability, though numerous literatures have emphasized the potential of technology to interact and significantly impact the three pillars of sustainability. Digital transformation was a major theme in all the literatures chosen for thematic analysis with emphasis on the need to adopt upcoming technologies like Blockchain. It was interesting to see that this call for adoption of

latest technologies was to prepare the industry to sustain through the different revolutions like Education 4.0.

For RQ2, based on the evaluation results, it can be concluded that the chosen algorithms are a good fit for the data chosen and that they are in line with the information summarized in the literatures discussed earlier in the report. THE-IR data worked well with traditional models like Logistic Regression, Random Forest. The data is simple to work with, easy to manage and can be analysed using widely used python libraries like pandas and NumPy. The data provides information on the current progress made in all SDGs but fails to provide a baseline on what that information means. For e.g.: an HEI with a score of 90 in SDG4 for 2024 and a score of 70 in the same SDG for 2025 cannot be categorized with certainty as

- The reduction in score was due to outcomes of efforts/strategies employed in year 2024 was unsuccessful (or)
- The reduction in score was due to lack of conclusive evidence prompting a higher contribution(effort/strategy) towards that SDG by that HEI for that year.

In addition, THE-IR does not provide a detailed view into actual progress made by an HEI on the ground due to the possibility that the scores could fluctuate every year.

7. Conclusion and Future Work

To conclude, the analysis conducted for RQ1 to include technological dimension as the fourth pillar of sustainability did not provide necessary information and foundation to pursue the research. Digital Transformation was a major theme in all the literatures chosen for content analysis. More research and analysis are needed to understand if there is a possibility to proceed with this direction as this would bring more structure to the adoption and implementation of technology in the Higher Education industry.

Similarly, the analysis conducted for RQ2 was successful as the chosen algorithms were a good fit for the THE-IR datasets. The model metrics were commendable underlining the aptness of the chosen design to create the variables included in the analysis. The prominent presence of missing values in the data, gap in interpretation of the scores given for each SDG had a hindering effect on the ability to conclude if THE-IR data can be used solely in predict the sustainable outcomes of HEI's business strategy. Nevertheless, THE-IR data can be used for sustainability related analysis of HEIs, as underlined in different literatures, either in combination with other metrics or as a yearly benchmark to gauge the scenario for any given year. As a possible direction for future work, compatibility study of THE-IR data with other sustainability assessment tools/metrics can be pursued to understand how HEIs can fine tune their efforts/strategies toward achieving long-term sustainability.

References

Basheer, N., Ahmed, V., Bahroun, Z., & Anane, C. (2025). Sustainability assessment in higher education institutions: exploring indicators, stakeholder perceptions, and implementation challenges. *Discover Sustainability*, 6(1), 252. https://doi.org/10.1007/s43621-025-01116-w

Avelar, A. B. A., & Pajuelo-Moreno, M. L. (2024). *Role of Higher Education Institutions in Promoting Sustainable Development Goals Through Research, Teaching and Outreach* (pp. 557–578). https://doi.org/10.1007/978-3-031-65909-6_31

Deda, D., Tesch, L., Gerv.sio, H., & Quina, M. J. (2025*). Sustainability Assessment of Higher Education Institutions According to Times Higher Education Impact Ranking (pp. 1255–1270).* https://doi.org/10.1007/978-3-031-80434-2_68

Ordonez-Ponce, E., Khare, A., & Khare, K. (2024). Canadian HEIs' contribution to the SDGs: what do the times higher education impact rankings unveil? *International Journal of Sustainability in Higher Education.* https://doi.org/10.1108/IJSHE-03-2024-0188

Nunez-Naranjo, A. F., Morales-Molina, T., & Quesada, A. C. (2025). *Relationship Between Technology and Education for Sustainable Development in Latin America as a Goal of the 2030 Agenda (pp. 345–355).* https://doi.org/10.1007/978-3-031-93103-1_34

Tafese, M. B., & Kopp, E. (2025). Education for sustainable development: analyzing research trends in higher education for sustainable development goals through bibliometric analysis. *Discover Sustainability*, *6*(1), 51. https://doi.org/10.1007/s43621-024-00711-7

Nogueira, E., Gomes, S., & Lopes, J. M. (2025). Unveiling triple bottom line's influence on business performance. *Discover Sustainability*, *6*(1), 43. https://doi.org/10.1007/s43621-025-00804-x

Kusmendar, Asih, A. M. S., & Normasari, N. M. E. (2025). Exploring sustainable pathways: A systematic literature review of three pillars of sustainability applications. *Sustainable Futures*, *10*, 100970. https://doi.org/10.1016/j.sftr.2025.100970

Calderon, A. (2023). Sustainability Rankings: What they are About and How to make them Meaningful. *Journal of Studies in International Education*, *27*(4), 674–692. https://doi.org/10.1177/10283153231172022

Urbano, V. M., Arena, M., Azzone, G., & Mayeres, M. (2025). Sustainable development in higher education: An in-depth analysis of Times Higher Education Impact Rankings. *Journal of Cleaner Production*, *501*, 145302. https://doi.org/10.1016/j.jclepro.2025.145302

Shawe, R., Horan, W., Moles, R., & O'Regan, B. (2019). Mapping of sustainability policies and initiatives in higher education institutes. *Environmental Science & Policy*, *99*, 80–88. https://doi.org/10.1016/j.envsci.2019.04.015

Okur, Ö., Huang, M., Angeli, L., van der Voort, H., & Huang, Y. (2025). Sustainable digital education technologies: an analysis of selection processes in European universities. *Discover Sustainability*, *6*(1), 204. https://doi.org/10.1007/s43621-025-01008-z

Basheer, N., Ahmed, V., Bahroun, Z., & Anane, C. (2025). Sustainability assessment in higher education institutions: exploring indicators, stakeholder perceptions, and implementation challenges. *Discover Sustainability*, *6*(1), 252. https://doi.org/10.1007/s43621-025-01116-w

Correa-Peralta, M., Vinueza-Martínez, J., & Castillo-Heredia, L. (2025). Evolution, topics and relevant research methodologies in business intelligence and data analysis in the academic management of higher education institutions. A literature review. *Results in Engineering*, *25*, 103782. https://doi.org/10.1016/j.rineng.2024.103782

Cruz, M. M. P., & Lumauag, R. G. (2024). Comparative Analysis of Machine Learning Algorithms for Predicting Student Academic Performance in Higher Education. *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, 888–896. https://doi.org/10.1109/ICUIS64676.2024.10866086

Trevisan, L. V., Eustachio, J. H. P. P., Dias, B. G., Filho, W. L., & Pedrozo, E. Á. (2023). Digital transformation towards sustainability in higher education: state-of-the-art and future research insights. *Environment, Development and Sustainability*, *26*(2), 2789–2810. https://doi.org/10.1007/s10668-022-02874-7

Aria, M. & Cuccurullo, C. (2017) bibliometrix: An R-tool for comprehensive science mapping analysis, Journal of Informetrics, 11(4), pp 959-975, Elsevier.

Bibliometrixdownload. Bibliometrix.org. https://www.bibliometrix.org/home/index.php/download

ImpactRankings. timeshighereducation.com. https://www.timeshighereducation.com/impactrankings

*How food connects all the sdgs(2016 June 14).*stockholmresilience.org https://www.stockholmresilience.org/research/research-news/2016-06-14-how-food-connects-all-the-sdgs.html

*Choosing the right estimator.*scikit-learn.org https://scikit-learn.org/stable/machine_learning_map.html