A Decision Tree Approach to the Collection of Data for Sustainable Development Goals

Sonique Amrika Ramlal
Georgia Institute of Technology
Atlanta, USA
soniqueramlal@outlook.com

Patrick Hosein
The University of the West Indies St. Augustine
St.Augustine, Trin and Tobgo
patrick.hosein@uwi.edu

Abstract

Traditional survey methods often suffer from inefficiencies including lengthy questionnaires and respondent fatigue, limiting their effectiveness in capturing social concerns. This study aims to optimize social issue surveys by employing a decision tree approach, using the Sustainable Development Goals (SDGs) as a structured framework for adaptive questioning. The survey was developed in JotForm, beta-tested with 25 participants, and then mass-distributed, collecting responses over a two week period. Results demonstrate that the decision tree approach enhances survey efficiency by dynamically adjusting question pathways based on respondent choices. Pruning and expansion decisions were informed by response rates, with low-engagement leaves identified for removal and high-engagement leaves marked for further branching in future iterations. The study highlights the potential for real-time survey adaptation and presents an efficient framework for conducting targeted social issue assessments, improving both data quality and respondent experience.

CCS Concepts

• **General and reference** → **Surveys and overviews**; • **Information systems** → *Data analytics*; • **Computing methodologies** → **Classification and regression trees**; *Supervised learning*.

Keywords

Decision Tree, Survey Optimization, Branching and Pruning, Sustainable Development Goals

ACM Reference Format:
Sonique Amrika Ramlal and Patrick Hosein. 2025. A Decision Tree Approach to the Collection of Data for Sustainable Development Goals. In *The Sixth edition of the International Conference on Digital Age & Technological Advances for Sustainable Development (DATA 2025), May 07–09, 2025, Tangier, Morocco.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3747897.3747905

Social issues are unfavorable conditions that negatively affect the personal or social lives of individuals. When these issues impact a significant portion of society, they become serious social concerns requiring attention. Examples of such issues include inequality and violence [7]. However, categorizing social issues can be challenging due to their subjective nature. These issues have wide-ranging impacts, from economic strain to psychological distress, making it essential to address them to improve quality of life. Surveys are a valuable tool for gaining insights into such concerns, as they can capture the perspectives of large populations. This study leverages surveys to identify and prioritize the most pressing social issues affecting citizens.

In 2015, the United Nations established 17 Sustainable Development Goals (SDGs) aiming to foster global transformation through a collaborative framework[13]. These goals target critical challenges faced by humanity, focusing on resolving social, economic, and environmental issues. The SDGs provide a structured foundation for understanding and addressing global and local concerns. This study adopts the SDG framework to design surveys that align with local priorities, using it as a basis for exploring the most significant social challenges within the community.

This research specifically focuses on eight SDGs that are most relevant to local social concerns: No Poverty (SDG 1), Good Health and Well-Being (SDG 3), Quality Education (SDG 4), Clean Water and Sanitation (SDG 6), Affordable and Clean Energy (SDG 7), Decent Work and Economic Growth (SDG 8), Reduced Inequality (SDG 10), and Peace, Justice, and Strong Institutions (SDG 16) [13]. These SDGs were selected due to their direct impact and relevance to the challenges faced by the population under study.

A decision tree consists of a root node, branch nodes, and leaf nodes, where the root represents the starting point, branch nodes make intermediate decisions, and leaf nodes contain final outcomes[4]. Decision tree pruning is the process of removing unnecessary branches from a decision tree to improve its efficiency [5]. When applied to survey design, decision trees enhance the ability to explore multiple aspects of a main idea while streamlining the process for respondents. This approach reduces the need for participants to answer every question, focusing instead on relevant areas.

The decision tree structure employed in this research allows for adaptive questioning, ensuring that survey respondents only answer the most relevant questions. Figure 1 provides a simplified visualization of this approach, where a broad category (the root node) branches into major social issues (intermediate nodes) and ends on specific details of the chosen issue (leaves).

The objective of this study is to demonstrate the use of a decision tree approach in traditional surveys, specifically for identifying and understanding the most critical social issues faced by citizens. By focusing on the eight SDGs most relevant to the local context, the survey reflects unique challenges while maintaining alignment with broader development priorities.

DATA 2025, Tangier, Morocco

ACM ISBN 979-8-4007-1359-0/25/05
https://doi.org/10.1145/3747897.3747905

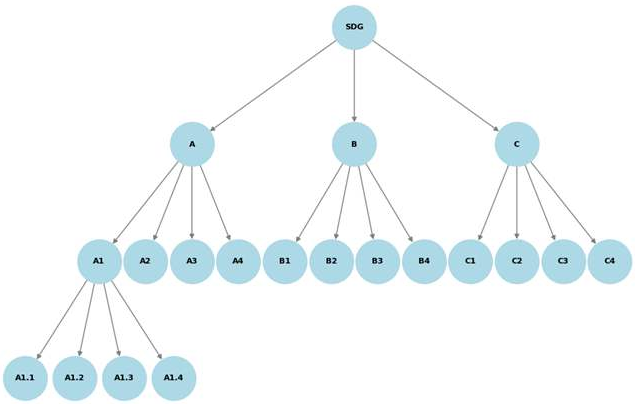

Figure 1: Decision Tree Branching and Expansion in Survey Design

1 Related Works and Contributions

1.1 Overview of Traditional Survey Methods

Traditional survey methods, such as paper-based surveys, have been historically integral to data collection. However, they often face challenges, including high costs, slow response times, and data quality issues. Emerging technology such as electronic and web-based surveys may serve as efficient alternatives to address these limitations. We begin by exploring the progression from traditional to optimized survey designs. Our goal is to provide insight into how advancements like decision tree based surveys can further refine data collection practices.

Boyer et al. [1] compared traditional print surveys with early computerized survey methods and found that computerized surveys, distributed via mailed computer disks, were easier for participants to complete. This improvement was largely attributed to the multimedia capabilities of electronic surveys. That is, these features may have contributed to clearer question presentation and richer open-ended responses. Despite these benefits, the manual distribution and collection process still incurred additional time and costs. In contrast, this research takes advantage of an online survey platform that automates data collection and integrates responses directly into a spreadsheet, significantly reducing time and eliminating manual errors by researchers.

Another critical aspect of survey design is data quality. Incomplete responses or inattentive behavior among participants are common issues. Boyer et al.[1] highlighted issues where respondents would either skip questions or rush through surveys, especially in cases involving lengthy questionnaires. They found instances where participants completed surveys at abnormally fast rates, suggesting inattentiveness. While their study relied on response time measurements to identify such cases, this research addresses the problem through decision tree logic. By dynamically presenting only relevant questions, the survey length is minimized [10]. This approach encourages more thoughtful and accurate responses by ensuring that participants do not encounter irrelevant questions.

Lonsdale et al.[8] compared paper-based and online survey formats and found that online surveys had fewer missing responses due to built-in prompts requiring participants to complete each item before progressing. Additionally, online surveys demonstrated faster response times while maintaining similar measurement reliability and validity to traditional paper surveys. While their study focused on comparing formats, this research optimizes the design of online surveys through sequential branching logic. The decision tree framework enhances adaptability by tailoring the survey path based on respondents' answers.

1.2 Previous Studies on Decision Tree Approaches

Decision tree models have been widely adopted across various fields to improve decision-making processes. Their ability to dynamically adapt based on user inputs has proven valuable in applications ranging from customer feedback collection to clinical assessments. Several studies demonstrate how decision trees can enhance data collection while addressing challenges such as redundancy, fatigue, and missing data.

Guerrero et al. [3] developed a computer adaptive survey (CAS) for measuring customer satisfaction in the telecommunications industry. It implemented a hierarchical decision tree framework that dynamically guides respondents through only the questions that are relevant to their concerns, reducing cognitive load and minimizing irrelevant questions. This approach not only improved response rates but also enhanced data quality by preventing disengagement. The weighting mechanism they implemented to reduce response bias provides an additional layer of optimization, though this feature is beyond the immediate scope of this research.

Jansen et al. [6] applied decision trees in clinical settings to optimize patient-reported outcome measures by reducing questionnaire length while maintaining measurement accuracy. Their study created an optimized version of the Boston Carpal Tunnel Questionnaire (BCTQ), reducing the number of items from 18 to a maximum of 6 without sacrificing accuracy by selecting only the most discriminatory questions based on answer patterns. Similarly, this research holds the shared goal of minimizing respondent burden while ensuring data accuracy. Tailoring question paths based on prior responses highlights the adaptability of decision tree methodologies in survey contexts.

Wehenkel et al. [14] further demonstrated the power of decision trees in optimizing classification tasks by applying them to power system security assessments. Their study showed that decision trees could efficiently identify potential stability risks without requiring extensive simulations. The key takeaway for this research is the importance of tree optimization, which ensures that decision paths are simplified without compromising accuracy. This concept directly informs the design of streamlined question pathways in surveys.

The study by Zheng et al. [15] emphasizes how decision trees can effectively process large-scale data while addressing challenges such as data sparsity and overfitting. In surveys, certain social issues may receive disproportionately low or high responses, similar to how rare crash events were under-represented in their study. By using decision tree based branching, survey structures can adapt to response distributions, ensuring that highly relevant issues are explored further while minimizing respondent fatigue.

1.3 SDGs in Addressing Social Issues

SDGs provide a broad framework for addressing global challenges, but not all SDGs are equally relevant in every context. A structured approach is essential to ensure that the most critical social issues are being highlighted in survey research. Decision trees offer a way to streamline this process by guiding respondents through relevant categories without overwhelming them with unnecessary questions.

One study analyzing global research trends related to the SDGs [11] found that some goals receive significantly more attention than others. This imbalance suggests that structured methodologies can help refine how SDGs are applied in research. In this research, eight SDGs were selected as the most relevant for capturing the key social issues affecting the country of study, ensuring data collection remains targeted.

An integrated framework for SDG implementation [2] further supports the need for structured methodologies. It emphasizes the importance of organizing SDG assessments into systematic categories, to facilitate targeted decision-making. Similarly, the decision tree approach in this research allows survey respondents to navigate through relevant topics while avoiding questions that do not apply to them.

Decision trees have also been used to improve sustainability assessments in the manufacturing sector [9]. A study applying decision tree based methodologies in manufacturing demonstrated how structured decision-making can improve data collection and guide organizations in prioritizing actions. The study's approach mirrors the objective of this research, which is to enhance survey methodologies by structuring responses in a way that ensures efficiency.

In retrospect, decision trees are valuable tools for optimizing data collection, improving engagement, and ensuring relevance in structured methodologies. By applying this framework to social issue surveys under the SDGs, this research builds on prior studies to develop a survey model that prioritizes relevance and reduces respondent burden.

2 Methodology

This research serves as a proof of concept, demonstrating the optimization of social issue surveys through a decision tree approach, utilizing the SDGs as a framework. It was designed to show how survey efficiency can be improved using adaptive question flows, while still providing a structured analysis of the most pressing social concerns. Consequently, the data collected was used to validate this approach rather than for in-depth insights, serving to inform future refinements to the survey structure and method.

Survey development began with the initial generation of questions using ChatGPT, followed by manual modifications to ensure relevance to the twin-island nation of Trinidad and Tobago. Questions were adapted to align with SDG targets while considering local, social and economic contexts. The survey was built using JotForm, leveraging its conditional logic capabilities to implement a decision tree format. This ensured that respondents were only presented with follow-up questions relevant to their chosen concerns. A beta test involving 25 participants was conducted over one week, allowing for adjustments in question clarity and survey flow. Following these refinements, the final survey comprised 129 multiple-choice questions, 4 demographic questions, and 1 open-ended response field.

The survey was mass-distributed via various digital channels, including LinkedIn, WhatsApp, Facebook, and TTLAB website. After a two-week data collection period, a total of 168 valid responses were gathered. JotForm was integrated with Google Sheets to facilitate real-time data storage and exportation to a CSV file for further analysis using Python. Data cleaning and preprocessing were conducted using Python. Exploratory data analysis was performed to examine demographic distributions and engagement trends.

Pruning and expansion decisions were based on a systematic response analysis. Leaves, or final follow-up questions, were evaluated for their engagement levels. Leaves with zero responses were marked for removal, as their absence would not affect data quality. Conversely, leaves with a response rate exceeding twice the average across all leaves of that SDG were considered for expansion by splitting them into two questions to capture more detail. To assess engagement levels of respondents and the effect on data quality, the open-ended question of the survey was manually classified into categories of 'meaningful' and 'not meaningful'.

Our survey was designed to accommodate repeated participation by the same respondents. To maintain accurate representation of the largest social issue pain points, a framework for weighting multiple responses was developed. The proposed weight used is $1/n$ where where n represented the number of times a respondent participated. Under this approach: A respondent with one submission would retain a full weight of 1, a respondent with two submissions would have each submission weighted 0.5 and a respondent with three submissions would have each submission weighted 0.33.

Respondent data from our survey revealed no duplicate submissions, meaning the weighting framework was not applied to this analysis. However, if this survey is conducted repeatedly over time, response weighting ensures fair representation of evolving social issues while minimizing bias from multiple responses and providing accurate trend analyses.

3 Results

3.1 Response Rates Across SDGs

First we examine the response rate at the first branch of our decision tree where participants were asked to select the social issue adapted from SDGs that is of most concern to them.

The most selected SDGs were 'Peace Justice and Strong Institutions' and 'Decent Work and Economic Growth', each accounting for 30.4 percent of total responses. The least selected SDGs were 'Affordable and Clean Energy', 'Inequality' and 'Quality Education', collectively receiving only 6.6 percent of responses.

3.2 Short-Answer Response Analysis

The final mandatory question in the survey prompted all respondents to elaborate on their concerns in an open-ended format. This analysis evaluates the qualitative engagement by categorizing responses as either 'meaningful' or 'not meaningful', where meaningful responses provided detailed insights, and non-meaningful responses were not informative including entries such as 'not applicable' or 'no'.

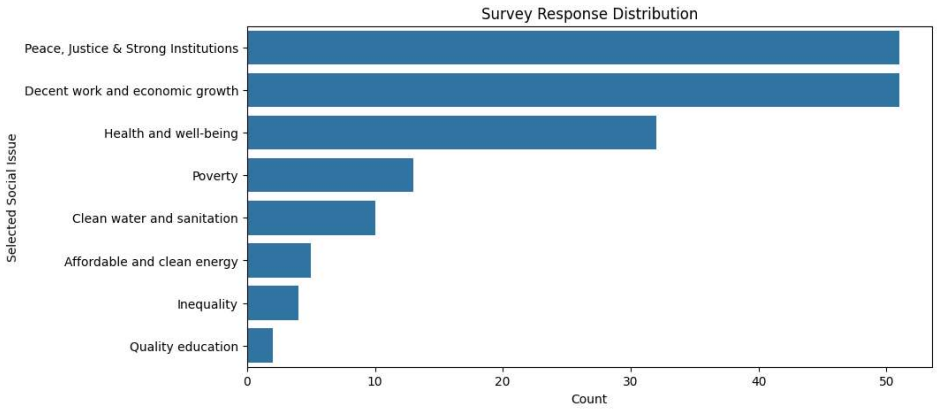

Figure 2: Survey Response Distribution Across SDGs

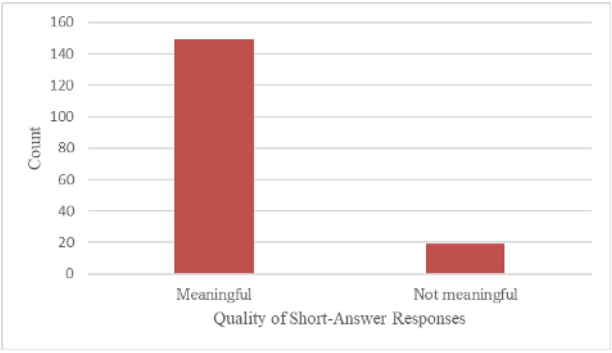

Figure 3: Short-Answer Response Quality

Figure 3 illustrates the distribution of response quality across the survey, showing a significantly higher proportion of meaningful responses compared to non-meaningful ones.

3.3 Branching and Pruning Analysis

Table 1 presents an example of the branching and pruning analysis conducted utilizing Peace, Justice and Strong Institutions. To determine necessary modifications for improving survey design as it relates to this SDG, the Leaf Question column represents the final follow-up questions for those respondents and the Response Rate was calculated as the percentage of respondents who answered the question compared to the total who reached the branch.

Table 2 lists the questions identified for expansion through further branching, along with their corresponding SDGs. It presents a broader perspective, highlighting the final follow-up questions from all SDGs that exceeded the expansion threshold, warranting further subdivision to capture more specific insights.

4 Discussion

4.1 Survey Refinement

Question generation using ChatGPT required continuous refinement through iterative prompting and manual adjustments. The generated questions did not initially align with the eight SDGs under consideration and were often overly specific. Instead, a decision tree structure with a clear progression from general to more specific issues per SDG was required to obtain maximum detail on respondents' most critical concerns. To adequately demonstrate the volume of data that could be collected, the survey was designed with four branching levels in addition to four demographic questions and concluded with an open-ended question for further elaboration. Although each respondent answered only nine questions in total, the full survey consisted of 134 questions spanning eight SDGs, illustrating the efficiency of the decision tree approach in gathering targeted data while minimizing survey fatigue.

Beta testing revealed several necessary modifications to improve clarity and efficiency. For example, the open-ended question initially asked, "Can you briefly describe your most recent experience where this social issue has impacted you? Include the approximate date of the incident." This wording implied that responses should only describe direct personal experiences. Instead, the objective was to gather broader perspectives on how respondents were affected. The question was therefore re-worded to "Can you describe the impact of this social issue on yourself or society? Please relay an experience if applicable." This modification ensured that respondents felt encouraged to provide relevant insights without being restricted to personal incidents.

4.2 The Role of Branching and Pruning

The decision tree survey structure was refined through a branching and pruning approach based on response rates at each level. Pruning decisions were made by removing leaves with 0 percent response rates as in table 1 with voter participation and awareness of rights and responsibilities. Conversely, leaves with more than twice the average response rate were identified for expansion. For example, community safety was selected for further breakdown into two additional questions with a response rate of 29.41 percent. Leaves with moderate response rates remained unchanged. This branching and pruning process ensures that future iterations of the survey removes under-utilized questions while enhancing engagement for high-priority issues. Removing unused leaves minimizes survey fatigue, while supporting a targeted data collection approach.

Table 1: Leaf Response Statistics for Peace, Justice and Strong Institutions

Leaf Question	Responses	Rate (%)	Action
What aspect of community safety affects you most?	15	29.41	Expand
What about violent crime affects you most?	8	15.69	No change
What about bribery and corruption affects you most?	5	9.80	No change
What about public trust in institutions affects you most?	5	9.80	No change
What about accountability of officials affects you most?	5	9.80	No change
What type of property crime affects you most?	4	7.84	No change
What about law enforcement response affects you most?	4	7.84	No change
What about community organizing affects you most?	3	5.88	No change
What about government transparency affects you most?	2	3.92	No change
What about voter participation affects you most?	0	0.00	Remove
What about public involvement in decision-making affects you most?	0	0.00	Remove
What about awareness of rights and responsibilities affects you most?	0	0.00	Remove

Table 2: Survey Questions Identified for Expansion Across SDGs

SDG	Question
1	What about job availability affects you most?
3	What about affordability of healthcare services affects you most?
3	What about nutrition and diet affects you most?
4	What about specialized skills courses or certification affects you most?
4	What about learning opportunities affects you most?
6	What about monitoring and regulation affects you most?
6	What about water distribution infrastructure affects you most?
6	What about seasonal water availability affects you most?
7	What about high upfront cost of renewable energy systems affects you most?
7	What about limited availability of renewable energy options affects you most?
8	What about job security affects you most?
8	What about wages and compensation affects you most?
10	What about unequal access to social services affects you most?
10	What about racial or ethnic discrimination affects you most?
10	What about discrimination based on sexual orientation affects you most?
16	What aspect of community safety affects you most?

4.3 Dynamic Survey Adaptation

While this research serves as a proof of concept, the decision tree approach has the potential for real-time adjustments when properly deployed. For example, data from our implementation identified Peace, Justice and Strong Institutions and Decent Work and Economic Growth as the most pressing social concerns. However, these may not consistently remain the most critical issues. Keeping the survey open for repeated participation can reveal changes in concerns. Moreover, the dynamic nature of this approach allows for adaptive branching and pruning in accordance to changing trends. This flexibility ensures that the survey remains responsive to evolving concerns while maintaining conciseness.

4.3.1 Open-ended Response Quality. Figure 3 suggests that most participants were willing to elaborate on their concerns, reinforcing the effectiveness of the open-ended question. Additionally, the high

level of engagement at this final stage of the survey indicates that respondents remained attentive throughout the survey process, which may suggest questionnaire fatigue was minimal. It is likely that this may be attributed to the decision tree approach which ensured only relevant follow-up questions were presented. However, non-informative answers account for 11 percent of responses, indicating that some respondents may have struggled with response structure or did not feel compelled to provide additional details. To improve future iterations of the survey, adjustments such as providing response examples or introducing a dropdown of general-concern elaborations for respondents to select from could be considered. These modifications may help reduce vague responses and increase the overall response quality.

5 Limitations

5.1 Phone Surveys and Large Language Model Integration

One limitation of online surveys is the potential exclusion of respondents without consistent internet access, which can introduce bias by under-representing those in remote areas. Implementing phone-based surveys as an alternative distribution method may produce a more representative sample of the population.

In this approach, researchers conduct direct phone calls, asking respondents survey questions in real time while following the decision tree structure. Just as in the online version, responses would determine the subsequent follow-up questions, allowing for a streamlined and relevant questioning process. Previous studies have explored the integration of Large Language Models (LLMs) to facilitate phone surveys, particularly in data analysis. LLMs have been used to process speech-to-text responses, automating the transcription of qualitative data [12]. This technology reduces the manual effort of phone surveys, ensuring efficient data processing. Future implementations could leverage similar techniques to refine qualitative analysis while maintaining the decision tree framework.

6 Future Work

The decision tree survey structure allows for continuous refinement. Future implementations could incorporate longitudinal studies, where the same respondents participate in the survey at multiple time points. Applying time-series analysis would enhance the ability to monitor social challenges. This approach would enable researchers to measure how social concerns shift over months or years. Additionally, we are currently collaborating with The Cropper Foundation, a civil society organization (CSO) active in Sustainable Development Goal (SDG) initiatives. This collaboration offers the potential for integrating our decision-tree survey tool within national SDG monitoring efforts.To further improve accessibility and usability, interactive dashboards could be developed to visualize survey results in real time.[12] These dashboards would allow researchers and policymakers to explore trends across different demographics and geographic regions, ensuring that data-driven insights are easily interpretable.

We also plan to implement the automated survey taking process presented in [12] to take into account people with only phone access. Future extensions could include the development of a mobile application to ensure broader accessibility, particularly in low-connectivity regions. Unlike web-based surveys that require a continuous internet connection, a mobile-based system would allow offline data collection by storing responses locally and syncing them when connectivity is available, thereby reducing infrastructure barriers to participation. LLMs could also be used to automate real-time response classification and summarization, especially for open-ended feedback. By using GPT-4 or similar models, qualitative responses can be categorized by concern type and summarized for policymakers or citizens. This enables the transformation of unstructured text into actionable insights with minimal manual intervention.[12]

References

[1] Kenneth K. Boyer, John R. Olson, Roger J. Calantone, and Eric C. Jackson. 2002. Print versus electronic surveys: a comparison of two data collection methodologies. *Journal of Operations Management* 20, 4 (2002), 357–373. doi:10.1016/S0272-6963(02)00004-9

[2] David Griggs, Mark Stafford Smith, Johan Rockström, Marcus Öhman, Owen Gaffney, Gisbert Glaser, Norichika Kanie, Ian Noble, Will Steffen, and Priya Shyamsundar. 2014. An integrated framework for Sustainable Development Goals. *ECOLOGY AND SOCIETY* 19 (12 2014), 49. doi:10.5751/ES-07082-190449

[3] Rafael Guerrero, Navin Dookeram, and Patrick Hosein. 2021. A Decision Tree Approach to Customer Surveys. In *2021 Second International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*. 74–81. doi:10.1109/IDSTA53674.2021.9660824

[4] IBM. 2021. what is a Decision Tree? https://www.ibm.com/think/topics/decision-trees Accessed: 2025-02-27.

[5] IBM. 2022. Pruning Decision Trees. https://www.ibm.com/docs/en/db2/11.1?topic=view-pruning-decision-trees Accessed: 2025-01-27.

[6] Miguel C Jansen, van der Oest Mark, Harm P Slijper, Jarry T Porsius, and Ruud W Selles. 2019. Item Reduction of the Boston Carpal Tunnel Questionnaire Using Decision Tree Modeling. *Archives of Physical Medicine and Rehabilitation* 100, 12 (June 2019), 2308–2313. doi:10.1016/j.apmr.2019.04.021

[7] Rebecca Kulik. 2025. Social Issue. https://www.britannica.com/topic/social-issue Accessed: 2025-02-14.

[8] Chris Lonsdale, Ken Hodge, and Elaine Hargreaves. 2006. Pixels vs. Paper: Comparing Online and Traditional Survey Methods in Sport Psychology. *J Sport Exerc Psychol* 28 (03 2006). doi:10.1123/jsep.28.1.100

[9] Justyna Patalas-Maliszewska, Hanna Łosyk, and Matthias Rehm. 2022. Decision-Tree Based Methodology Aid in Assessing the Sustainable Development of a Manufacturing Company. *Sustainability* 14 (05 2022), 6362. doi:10.3390/su14106362

[10] Sindre Rolstad, John Adler, and Anna Ryden. 2011. Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-analysis. *Value in Health* DOI: 10.1016/j.jval.2011.06.003 (12 2011). doi:10.1016/j.jval.2011.06.003

[11] Amanda Salvia, Walter Filho, Luciana Brandli, and Juliane Griebeler. 2018. Assessing research trends related to Sustainable Development Goals: local and global issues. *Journal of Cleaner Production* 208 (10 2018). doi:10.1016/j.jclepro.2018.09.242

[12] Trevon Tewari and Patrick Hosein. 2024. *Automating the Conducting of Surveys Using Large Language Models.* IEEE, 136–151. doi:10.1007/978-3-031-66705-3_9

[13] United Nations. 2015. The 17 Goals: Sustainable Development. https://sdgs.un.org/goals Accessed: 2025-01-27.

[14] L. Wehenkel and M. Pavella. 1993. Decision tree approach to power systems security assessment. *International Journal of Electrical Power & Energy Systems* 15, 1 (1993), 13–36. doi:10.1016/0142-0615(93)90014-E

[15] Zijian Zheng, Pan Lu, and Denver Tolliver. 2016. Decision Tree Approach to Accident Prediction for Highway–Rail Grade Crossings: Empirical Analysis. *Transportation Research Record: Journal of the Transportation Research Board* 2545 (01 2016), 115–122. doi:10.3141/2545-12