AbstractA Machine Learning Approach to Predicting Lifestyle Sustainability Scores Using Behavioral and Environmental Factors
By Sanjana Shamarayar, Mentor: Abdulla Kerimov

Quantifying sustainability scores using lifestyle and environmental attributes provides a personalized evaluation to promote behavioral change. This study evaluates the effectiveness of ensemble-based modeling methods in predicting individual sustainability scores from complex data patterns. Four machine learning algorithms—Random Forest, Extra Trees, LightGBM, and XGBoost—were implemented, incorporating hyperparameter tuning, bootstrapped testing, lumping approaches, cross-validation, and grid search to optimize performance. All models demonstrated strong predictive ability with minimized overfitting and solid generalization. The 3-class XGBoost model achieved the highest predictive accuracy, with bootstrapped F1 scores of 0.77 [CI 0.67–0.87] for Moderate Sustainability and 0.84 [CI 0.75–0.93] for High Sustainability, outperforming other 3-class models despite wider confidence intervals. SHAP analysis identified Transportation Mode, Environmental Awareness, Monthly Water Consumption, Monthly Electricity Consumption, Disposal Methods, and Home Size as the most influential predictors, enhancing interpretability and applicability across diverse populations. However, reliance on self-reported survey data, a relatively small blind-testing dataset, and a limited number of input features remain key limitations that may restrict generalizability, highlighting important considerations for future research.

Keywords: sustainability, lifestyle, environmental, demographic, factors, machine learning, feature importance analysis, SHAP, hyperparameter tuning, bootstrapped testing

Introduction

In a world where environmental degradation and resource depletion are growing global concerns, increasing individual sustainability is essential for planetary well-being. Although environmental sustainability is often addressed through preventative laws and large-scale measures, little action occurs at the personal level (Todi). An individual's lifestyle behaviors—often overlooked in sustainability promotion efforts—accurately reflect their environmental impact (Farhud). As environmental awareness expands, the need to evaluate sustainability behaviors to encourage eco-friendly and responsible practices grows as well. Recognizing individual sustainability levels is therefore a critical step toward achieving broader environmental awareness goals.

While multiple methods exist to evaluate sustainability, some are more effective than others. A qualitative approach is necessary to assess the wide range of interconnected factors influencing sustainability, yet this complexity makes global comparisons challenging (Farhud). Sustainability cannot be determined by a single action but rather by patterns of behavior over time, requiring integration of multiple variables that differ in impact, size, importance, and reliability. Traditional methods—such as surveys, questionnaires, and point-based scoring

systems—are often detailed but may fail to account for regional differences and global environmental variation (Mansour). For example, point-based systems assign fixed values per response without considering correlations between variables, limiting their ability to capture complex relationships and potentially introducing human bias that may mislead sustainability efforts.

To avoid these oversimplifications, sustainability prediction methods should not assume linear relationships and must account for adaptability among variables. Scoring logic should be globally applicable, accurate, and relevant, minimizing inaccuracies to better reflect real-world sustainability trends. Recent advances in machine learning (ML) offer a promising solution. ML models can analyze large, versatile datasets and identify complex, multidimensional relationships between lifestyle behaviors and sustainability outcomes that traditional surveys may miss. By generating more personalized, data-driven predictions, ML-based approaches can provide meaningful feedback to support environmental research, education, and behavioral change. As sustainability assessment becomes increasingly data-driven, examining complex lifestyle interactions can encourage more informed action toward a greener future.

Relevant Works

The use of ML to create sustainability prediction models have been implemented among specific participant groups, such as college students, as researched by Wang et al. (2022). The predictive power of machine learning, specifically a decision-tree model, was observed in this study, by forecasting pro-environmental behavior among students attending college. In this study, 336 university students in Guangdong Province, China were given a survey that measured pro-environmental factors including environmental activism, social identity, and willingness to behave environmentally-responsible. A decision-tree model was implemented to find the factors with the highest correlation with pro-environmental behavior among college students. The results showed that the strongest predictors of pro-environmental behavior include having a willingness to behave in an environmentally responsible manner, perceived behavioral control, and innovative behavior. Additionally, Wang et al. (2022) concluded that using a machine learning model like decision-tree can expand the limited perspective of viewing pro-environmental behaviorisms, leading to further research on this prevalent topic.

Expanding beyond the integration of ML techniques for prediction models, Wang (2025) provides a comprehensive study on the implications of using ML to strengthen Life Cycle Assessment (LCA). LCA, a methodology used to quantify environmental impacts, has four phases that ML can be integrated into, in order to limit data gaps and timeliness: goal & scope, life cycle inventory, life cycle impact assessment, and interpretation. This study uses a bibliometric analysis and literature review to highlight the importance of ML techniques such as natural language processing (NLP), in terms of making LCA more progressive and transparent. Wang concludes that using ML offers significant improvements for sustainability assessments by filling in data gaps, conducting a predictive analysis, and limiting uncertainty within the assessment, underscoring the rising influence of ML in evaluating sustainability in research.

Similar to the decision-tree model used by Wang et al. (2022), Baehr et al. (2024) explores the utilization of ML in predicting life cycle environmental impacts on products by addressing challenges in the usage of traditional Life Cycle Assessment (LCA) formats. A large environmental product declarations (EPD) digitized dataset was used and analyzed with an artificial neural networks (ANN) model. The ANN model was used to estimate the environmental impact indicators among the wide range of products used in the dataset. Additionally, an in-depth uncertainty and sensitivity analysis was employed with residual Gaussian Process Regression (rGPR). Therefore, the hybrid ANN-rGPR model was used to reduce uncertainty within the model as well as increase predictive performance leading to optimal results. The results of this study concluded that ML can be effectively used to increase environmental impact estimation accuracy, although limitations still remain.

Research Gap and Contribution

Throughout the existing studies, many significant limitations persist in terms of specifically addressing the application of ML to sustainability level assessments. First, there are a very limited number of studies that focus on this domain, with most studies consisting of literature reviews that aren't enough to expand the scope of current research. In the range of these limited studies, most employ simple ML models that fail to achieve a strong predictive accuracy. This, in turn, limits the practicality of these models which can reduce its usability. Despite certain models resulting in a reasonable predictive performance, a lack of transparency and interpretability can create difficulties in applying the study to a larger scale. Additionally, a lack of scrupulous evaluation can pose issues with the reliability of the model as a whole. Blind testing and other heavily robust validation frameworks are often absent in these models which can raise concern regarding the generalizability of the model. Lastly, data quality issues including missing values, inconsistent data, and other measurement related uncertainties further undermine the reliability of pre-existing models. Collectively, these limitations underline the need to address data diversity and real world applicability in ML models.

In this study, the objective is to address these research gaps by developing a machine learning model that predicts individual sustainability scores using a self-collected, and publicly available dataset. By increasing the scope of participant diversity, this study aims to evaluate how demographic and behavioral factors influence sustainability and environmental impact. Using a standardized methodology and interpretable model will contribute to accurately identifying which lifestyle factors most strongly impact sustainability as a whole. In essence, this work provides valuable insight to contribute to the growing numbers of sustainability research by showing how ML can be applied on a larger scale and what factors affect model predictions.

Dataset

Our first dataset uses demographic information, living preferences, environmental impact indicators, and additional lifestyle factors to assign each of 499 participants a sustainability score on a numerical scale from 1 to 5 (Sharma). The dataset includes 19 lifestyle-based variables. Of

the 499 participants, 97 received a rating of 1, 34 received 2, 101 received 3, 91 received 4, and 176 received 5. The purpose of this dataset is to quantify an individual's sustainability level and environmental impact based on their lifestyle attributes. Variables include age, gender, residential setting (rural, suburban, urban), diet type, frequency of consuming locally sourced food, energy usage, transportation habits, home type and size, clothing purchase frequency, preference for sustainable brands, physical activity, community engagement, environmental awareness, electricity and water consumption, plastic usage, and waste disposal practices.

To test the optimal sustainability prediction model, a second survey-based dataset of 87 participants (approximately 17% of the original population) from the United States, Canada, and India was collected using a 20-question survey aligned with the same variables. Among them, 10 self-identified as "Low Sustainability," 73 as "Moderate Sustainability," and 4 as "High Sustainability." Participants self-assigned their sustainability level using both a numerical scale (1–5) and a categorical classification (low, moderate, high). While the diverse geographic locations and age range provide broader perspectives, the small sample size, limited circle of friends and family, and lack of random selection restrict generalizability and may limit the model's ability to capture the full diversity of backgrounds and behaviors in a larger population. The goal of this dataset is to compare participants' perceived sustainability ratings with the model-predicted sustainability scores based on their stated lifestyle factors.

Methodology

Data Preprocessing

The dataset used for this study required the encoding of categorical based data variables. These variables were encoded using one of two encoding methods: one-hot or ordinal encoding. One-hot encoding was utilized for categorical variables that had no inherent ordering. This includes Location (Urban, Suburban, Rural), Diet Type (Mostly Plant-Based, Balanced, Mostly Animal-Based), Transportation Mode (Bike, Public Transit, Car, Walk), Energy Source (Renewable, Mixed, Non-Renewable), Home Type (Apartment, House), Gender (Male, Female, Non-Binary), Disposal Methods (Composting, Recycling, Landfill, Combination), and Sustainable Brands (True, False).

However, variables that had a meaningful progression or ranking were ordinally encoded. These include Local Food Frequency, Clothing Frequency, Using Plastic Products, Community Involvement, and Physical Activities. Due to the level of frequency or engagement that had to be ranked, numbers on a scale of 1-3 were assigned to the specific variables correlating to the descriptions. Specifically, the frequency-related variables were encoded as Rarely = 1, Sometimes = 2, Often = 3, while the engagement-related variables were encoded as Low = 1, Moderate = 2, and High = 3.

The original dataset incorporated a sustainability rating from a scale of 1-5 assigned to the participant, given their lifestyle behavioral characteristics. Additionally, we explored an alternative lumping strategy that combines sustainability classification variables to maximize

class separability and balance the data. Based on the support (actual occurrence) values tested, ratings 1 and 2 were mapped to resemble a “Low sustainability”, ratings 3 and 4 “Moderate sustainability”, and 5 “High sustainability”.

Models and Evaluation Metrics

To evaluate the model, the dataset was divided into an 80/20 stratified train–test split. Stratification ensured balanced class distributions in both sets that accurately reflected the full dataset, preventing biases from disproportionate class representation and supporting reliable model evaluation (Sivakumar). The survey-based blind dataset was used to assess real-world generalization and verify consistency with train/test results. This independent dataset was collected to better reflect real-world conditions and improve model robustness through practical applicability. By accounting for natural irregularities often absent in curated datasets, the survey data strengthens generalizability.

The prediction framework evaluated Logistic Regression, KNN, Random Forest (RF), Extra Trees (ET), XGBoost (XGB), and LightGBM (LGBM), selected for usability and predictive accuracy. Random Forest combines multiple decision trees using bootstrap sampling and random feature selection to improve robustness and interpretability (Breiman). Extra Trees builds highly randomized trees with specific hyperparameter tuning to enhance efficiency; while reducing variance, it may introduce greater bias (Geurts). XGBoost and LightGBM are gradient boosting frameworks widely used for regression and classification. XGBoost applies tree boosting to enhance scalability and performance, producing state-of-the-art results while reducing overfitting compared to traditional gradient boosting methods (Chen). LightGBM utilizes histogram-based learning and a leaf-wise tree growth strategy, offering high efficiency and speed while maintaining comparable accuracy (Ke).

Model performance was assessed using precision, recall, and F1 score. Precision represents the proportion of non-false alarms, or correctly predicted outcomes over all predicted outcomes, and was compared across all train/test split models to identify the highest values with the fewest false alarms (Conciatori). Recall measures the proportion of correctly predicted outcomes over all actual outcomes, reflecting sensitivity and overall model effectiveness. The F1 score, the harmonic mean of precision and recall, provides a balanced summary of model quality and serves as an optimal metric for comparing performance across the four models.

Hyperparameter Tuning

To optimize predictive performance, hyperparameter tuning was conducted using the Grid Search method with 5-fold cross-validation. Hyperparameter tuning systematically selects the most optimal hyperparameter values to ensure strong model performance (Arnold). Because model accuracy after splitting data into training and testing sets depends on the chosen hyperparameters, Grid Search efficiently evaluated combinations across validation folds to reduce overfitting and maximize predictive performance (Arnold). Of the six original models,

four—Random Forest, Extra Trees, XGBoost, and LightGBM—were tuned, while Logistic Regression and KNN were treated as baseline models due to their performance results.

For Random Forest and Extra Trees, n_estimators, max_features, min_samples_split, and class_weight were tuned (Table 1). n_estimators (100–500) controls the number of trees; although more trees can improve performance, they may increase training time and risk overfitting (Kushan). max_features, varied from 1 to 36 in increments of 5, controlled features per split to promote diversity, reduce overfitting, and improve robustness (Kushan). min_samples_split (10–50, increments of 5) ensured meaningful data per split and reduced bias and overfitting by avoiding extreme values (Quadri). class_weight was set to "balanced" to address dataset imbalance and prevent bias toward majority classes (Chen, Chao).

For XGBoost and LightGBM, n_estimators (100–500), learning_rate, subsample, and colsample_bynode were tuned. learning_rate controls each tree's contribution through step size and must be tuned with n_estimators, as lower rates require more trees to reach similar performance; rates of [0.001, 0.002, 0.003, 0.01, 0.02, 0.03, 0.1, 0.2, 0.3] were tested for stability (Tayebi). subsample—the fraction of randomly selected samples per tree—was tuned using [0.5–1.0] to improve robustness and reduce overfitting (Tayebi; Sibindi). colsample_bynode, the fraction of features per node split, was also tuned within [0.5 - 1.0] to enhance randomness and generalizability without increasing bias or overfitting (Gera).

Table 1. Model hyperparameters and their value ranges

Model	**Parameter**	**Value**
Random Forest	n_estimators	[100,500]
	max_features	between 1 and 36 with increment of 5
	min_samples_split	between 10 and 50 with increment of 5
	class_weight	"balanced"
Extra Trees	n_estimators	[100,500]
	max_features	between 1 and 36 with increment of 5
	min_samples_split	between 10 and 50 with increment of 5
	class_weight	"balanced"
XGBoost	learning_rate	[0.001, 0.002, 0.003, 0.01, 0.02, 0.03, 0.1, 0.2, 0.3]
	subsample	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
	colsample_bynode	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
	n_estimators	[100,500]

LightGBM	learning_rate	[0.001, 0.002, 0.003, 0.01, 0.02, 0.03, 0.1, 0.2, 0.3]
	subsample	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
	colsample_bynode	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
	n_estimators	[100,500]

The training process of the model was evaluated using a 5-fold Cross Validation (CV) to provide a reliable estimate of the model's performance. In order to quantify and estimate the uncertainty, 95% confidence intervals were calculated using the CV distribution scores and values. Evaluating the confidence intervals between each of the models can be used to indicate the stability and fluctuations between the models' performances. This offers valuable insight into ranking and comparing the models in terms of which confidence intervals allow for more precision, versus more variability (Zhang, Jesse, et al).

Additionally, a bootstrap method was used to assess the model performance on a test set. Given the sample size, 1000 bootstrap samples were generated through the process of resampling of test data (Zhang, Jesse, et al). From the performance outcomes calculated through this approach, the 95% confidence intervals can be computed and used to provide a more detailed understanding of the model's expected performance and uncertainty (Tsamardinos). This resampling-based evaluation method is crucial for robustly estimating uncertainty in performance metrics and enables a more reliable interpretation of model performance results.

Feature Importance Analysis

Feature importance analysis is a key tool for improving interpretability in machine learning models, as it quantifies the contribution of each input feature. This is especially important for complex models where variable relationships are not directly observable. By revealing the visible impact of each feature, it enables users to identify relevant predictors and make informed decisions (Zhao). It can also reduce model complexity by focusing on a smaller subset of highly influential features rather than using all 18 survey variables to predict sustainability scores. This improves efficiency, enhances interpretability, and allows closer examination of the most meaningful behavioral and environmental factors, ultimately helping to optimize overall performance. SHAP (SHapley Additive exPlanations) analysis is a widely used feature importance method rooted in cooperative game theory that provides both global and local interpretability (24). SHAP assigns each input feature a value based on Shapley values, representing its average marginal contribution across all feature combinations. The framework satisfies key properties of accuracy, consistency, and additivity, making it a robust and reliable approach for explaining model predictions (Wang, Huanjing, et al.).

Results and Discussion

This section presents a comprehensive evaluation of machine learning model performance for sustainability score prediction under both 5-class and 3-class frameworks. Ensemble-based models are compared using testing set F1 and recall scores, with class-wise performance differences examined alongside the effects of class imbalance, confidence interval variability, and model complexity. The reduced 3-class framework is evaluated to assess trade-offs between interpretability and accuracy, followed by a SHAP-based feature importance analysis to identify key predictors and assess model sensitivity.

For the 5-class models, Figure 1 compares test-set F1 scores across Logistic Regression, KNN, LightGBM, Random Forest, XGBoost, and Extra Trees. Ensemble models consistently outperformed baseline models for higher-frequency ratings. Rating 1 achieved the strongest performance across all models, with F1 scores ranging from 0.84–0.92, while Rating 2 showed the weakest performance (0.18–0.44), largely due to class imbalance, with baseline models ranging from 0.18–0.43 (Table A1). Ratings 3, 4, and 5 demonstrated intermediate to strong performance in ensemble models but lower results in baseline models. Logistic Regression and KNN achieved F1 scores of 0.84 and 0.92 for Rating 1, respectively (Table A1), but declined for Ratings 3 and 5, and substantially for Ratings 2 and 4. Logistic Regression's lowest F1 was 0.22 for Rating 4, while KNN's lowest was 0.18 for Rating 2, reflecting limited ability to capture non-linear relationships.

Among ensemble models, XGBoost achieved the strongest overall performance, with bootstrapped F1 scores of 0.92 CI [0.82–1.00] for Rating 1, 0.77 CI [0.65–0.87] for Rating 3, and 0.84 CI [0.77–0.92] for Rating 5 (Table A2), demonstrating high predictive ability and relatively tight confidence intervals indicating lower variance and strong generalizability. LightGBM also performed strongly, achieving 0.95 CI [0.86–1.00] for Rating 1 and 0.83 CI [0.75–0.91] for Rating 5 (Table A2). Higher-frequency ratings (Ratings 1 and 5 with 20 and 35 samples, respectively) showed narrower confidence intervals than lower-frequency ratings, a pattern also observed in Random Forest and Extra Trees. For example, Rating 5 in the Extra Trees bootstrapped test achieved 0.76 CI [0.70–0.82], corresponding with its higher sample size of 35. Confidence intervals were consistently narrower in CV tests than in bootstrapped testing, reflecting greater variability during generalization. Expanded intervals were especially evident in underrepresented ratings, such as Rating 2 in the RF bootstrapped test with CI [0.00–0.60] compared to the CV train CI [0.00–0.33] (Table A2), particularly affecting Ratings 2 and 4 and indicating increased uncertainty in these classes.

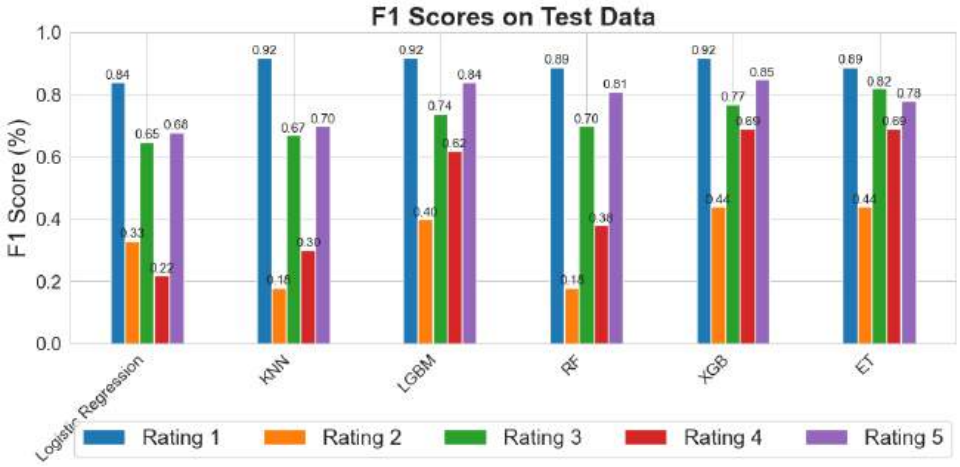

Fig 1. Test F1 scores to predict the 5 ratings of sustainability score by baseline logistic regression and KNN models and optimal tuned ensemble models

A 3-class model (with sustainability rating outputs of "Low", "Moderate", and "High") was implemented to better capture meaningful patterns in sustainability performance. The usage of this lumping method can reduce class imbalances present between ratings, reflecting a more nuanced assessment of the model's predictability. Figure 2 presents the test-data F1 Scores for predicting a 3-class sustainability rating across all ensemble models. Throughout all models, F1 scores varied among the three classes. Excluding RF, most models had the lowest F1 scores for Rating 2 (Moderate Sustainability) with a range of [0.67 - 0.73]. Rating 1 (Low Sustainability) had the highest F1 scores for Logistic Regression and XGB with scores of 0.81 and 0.86 respectively, reflecting a stronger model performance for this class among these models (Fig 2). Rating 3 (High Sustainability) had the highest F1 scores for KNN, LGBM, and ET, with scores of 0.69, 0.78, and 0.81 respectively; likewise, a stronger model performance is suggested in these ensemble models among this rating.

Similar to the 5-class model, baseline models KNN and Logistic Regression had lower F1 Scores in comparison to the more complex ensemble ML models. Between Logistic Regression and KNN, Logistic Regression achieved the highest test F1 scores for each of the rating categories: Rating 1: 0.81, Rating 2: 0.68, Rating 3: 0.78. KNN however, demonstrated a lower performance with scores of 0.68, 0.67, and 0.69 for Ratings 1, 2, and 3 (Table A3). On the other hand, the tuned models presented in Table A4 typically outperformed their baseline counterparts, with XGBoost achieving the strongest test F1 scores.

XGBoost achieved the highest bootstrapped test F1 scores with values of 0.82 CI [0.71 - 0.93] for Rating 1, 0.77 CI [0.67 - 0.87] for Rating 2, and 0.84 CI [0.75 - 0.93] for Rating 3. These high-performing evaluation metrics are reflective of the model's ability to capture non-linear relationships within the data. Additionally, the relatively narrow confidence intervals indicate robust and stable performance despite variability introduced through the bootstrapping procedure. LightGBM and Random Forest followed closely behind with moderate to high achieving F1 scores. While Random Forest achieved a high F1 score of 0.83 CI [0.70 - 0.94] for

Rating 1 in the bootstrapped test set, LightGBM demonstrated strong performance for Rating 3 with an F1 score of 0.81 CI [0.72 - 0.89] (Table A4). These results suggest that individual models may excel for specific classes; however, their performance was less consistently strong across all ratings compared to XGBoost. Extra Trees exhibited mid-range F1 scores for all three rating categories as seen in Figure 2 (Rating 1: 0.72 CI [0.57 - 0.87], Rating 2: 0.75 CI [0.65 - 0.84], Rating 3: 0.74 CI [0.64 - 0.84]), indicating reduced effectiveness in capturing class-specific patterns.

Similar to the 5-class models, the 3-class models suggest narrower CI intervals in cross-validation (CV) training sets compared to the bootstrapped test sets. As observed in XGBoost, the Rating 1 CI in the CV training set was [0.74 - 0.83], in comparison to the bootstrapped test-set CI of [0.71 - 0.93]. The ± 0.05 CI width in the training set compared to the ± 0.11 CI width in the testing set illustrates the expected widening of confidence intervals during evaluation on unseen data. The increase in CI width reflects greater variance in test-set performance, contributing to class-specific uncertainty across certain models. Overall, these findings reinforce XGBoost as the most reliable and generalizable model for 3-class classification, making it well suited for SHAP-based interpretability analyses.

Comparing model performance between the 3-class and 5-class models reveals trade-offs between generalization, predictive ability, and interpretability. Although reducing the number of classes does simplify the prediction task, a probable tradeoff would be model performance across various ML algorithms and evaluation settings. For the baseline models, training performance typically improved when shifting from a 5-class to 3-class model. In terms of the Logistic Regression model, the Rating 1 F1 score increased from 0.82 to 0.86 between the two classes, although this improvement was not observed in the test-set (Tables A1, A3). The test-set exhibited the opposite pattern, as performance decreased from 0.84 in the 5-class model to 0.81 in the 3-class model. This suggests that while consolidating the classes simplified the training boundaries, it may have reduced class separability during testing.

Ensemble models on the other hand, demonstrated an overall stronger performance for 5-class models than the 3-class models for Rating 1. The 5-class XGBoost achieved a bootstrapped test F1 score of 0.92 CI [0.82 - 1.00], exceeding the 3-class test score of 0.82 CI [0.72 - 0.93] (Tables A2, A4). Similarly, the CV training performance was higher in the 5-class model (0.84 CI [0.74 - 0.93]) than the 3-class model (0.78 CI [0.74 - 0.83]). Although these results do indicate that XGBoost retained its high performance in a more complex 5-class setting, the 3-class model's narrow CV training CIs reflect improved stability during training. Despite the cost of a lower predictive performance on the test set, the reduced variation acts as a trade-off for this issue.

Both 3-class and 5-class modeling carry several pros and cons that must be considered before choosing the most optimal predictive model. The 3-class framework offers plenty of advantages in terms of interpretability and an improved class balance, making it suitable for feature importance analysis. However, this simplification can negatively impact the predictive performance for certain ratings. Contrastingly, the 5-class framework enables a deeper look on

rating-specific details, especially when using high-capacity ensemble models. Nonetheless, the wider CIs for most 5-class algorithms may lead to increased variability and uncertainties within the model. Taking these implications into account, XGBoost consistently outperforms the baseline and other ensemble models across both the 3-class and 5-class models. It has achieved the highest F1 scores and maintained robust generalization as seen in the strong bootstrapped test performance and narrow confidence intervals. Considering the cons of a relatively lower predictive performance, the improved interpretability and stability provided by XGBoost modeling makes it suitable for future feature and SHAP analysis steps.

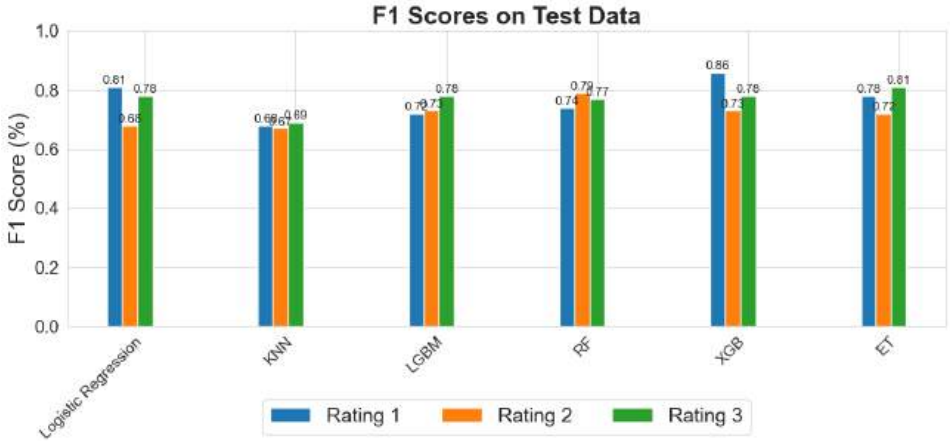

Fig 2. Test F1 scores to predict the 3 ratings of sustainability score by tuned optimal models

To further identify the most influential predictors of a sustainability score, feature importance analysis was conducted for the 3-class model. Specifically, SHAP Analysis was applied to the XGBoost Model, due to its high performance for feature interpretation. The SHAP Analysis results revealed the most predictive features of the model. We utilized the top 6 features with a CI range of [0.08 - 0.13] that exhibited the largest influence on model predictions (Fig 3). This includes: Transportation Mode, Environmental Awareness, Monthly Water Consumption, Monthly Electricity Consumption, Disposal Methods, and Home Size. Focusing on the most important features in the 3-class model allows for a clearer interpretability by reducing the noise from low-impact features. Additionally, limiting the analysis to the top features can enable more meaningful insights towards the primary influencers of the model's predictions, and increase performance optimization.

As shown in Figure 3, the mean absolute SHAP ranking plot orders the input features based on their overall contribution to the sustainability score. The more influential features including TransportationMode_Car and EnvironmentalAwareness have confidence intervals of 0.13 due to the higher variability as a result of higher SHAP magnitudes. These values are followed closely with MonthlyWaterConsumption (CI 0.11) and MonthlyElectricityConsumption (CI 0.10) which have a notable impact as well (Fig 3). This trend continues until a steep dip is noticed between UsingPlasticProdcuts (CI 0.07) and CommunityInvolvement (CI 0.04),

indicating the clear separation between high-impact and low-impact features. These results visually illustrate that daily-lifestyle and consumption based factors have a greater predictive power than demographic factors in the 3-class framework. Demographic characterizations such as Age, Gender, and Location have a minimal impact as seen in the low mean SHAP values in Figure 3. These findings contribute to the overall claim that sustainability-related lifestyle behaviors including resource-consumption factors are the dominant factors that influence the model's output.

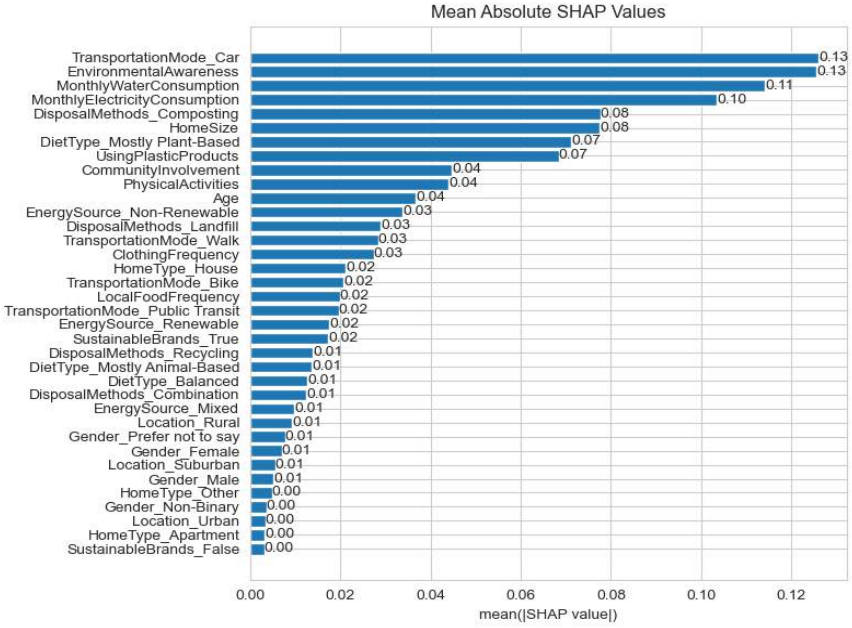

Fig 3. Mean Absolute SHAP Ranking Plot

Figure 4 illustrates the sensitivity of the test macro F1 score to the number of highest ranked features that were chosen using SHAP analysis. The macro F1 score, an evaluation metric that assesses the average F1 scores across all classes, was used after hyperparameter tuning the model. This reveals the top features that were most influential in predicting the sustainability score. As seen in the figure, the model performance increases as the number of important features increase from 2 - 6. This shows an improvement from 0.63 (2 features) to 0.77 (6 features) in the test macro F1 score value (Fig 4). The trend suggests that incorporating more high-importance features can allow the model to capture more relevant information, therefore improving its predictive performance.

Yet, after the number of features increases beyond 6 going to 8, the test macro F1 score slightly decreases to 0.76. A visible plateau and slight marginal drop is seen at this point, suggesting that adding more features beyond 6 may prevent the improvement of predictive performance within the model. All in all, the curved relationship between the number of features and the macro F1 scores indicate that the model performance is indeed sensitive to feature selection. Figure 4 illustrates that a moderate number of highly informative features yields the best predictive performance overall.

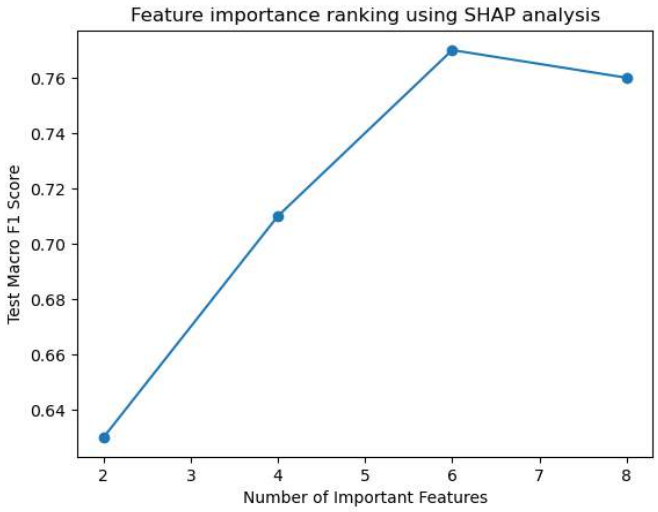

Fig 4. Test macro F1 score to predict 3 ratings of sustainability scores using top important features from SHAP analysis and tuned XGBoost model

The SHAP-based feature importance analysis reveals that not all features need to be incorporated into the model for it to encompass a high predictive performance. Instead, a strong performance can be achieved by using only an important subset of the original features rather than relying on all available inputs. This model reaches its highest test macro F1 score with 6 top-ranked features, highlighting the effectiveness of SHAP in observing the most influential predictors of sustainability scores (Fig 4). Accordingly, relying on a small subset of features can improve the model's interpretability while concurrently reducing the computational complexity and overfitting of the model.

In comparison to other models trained with a greater amount of features, the 6-feature framework demonstrates a substantial improvement in the model's class-balance as reflected by the macro F1 scores. Previous studies often prioritize maximizing accuracy by using a large number of input features, which can often impact interpretability. This study, in contrast, demonstrates that while predictive performance must be optimized in the model, it can still be achieved using a reduced feature set. Focusing on the macro F1 score, or average of the test scores, rather than the accuracy alone can provide a more balanced evaluation of the performance across various classes. This is often overlooked in related literature works.

Now, we present the results of the XGBoost 3-class model on blind test collected data under optimal conditions. The 3-class XGBoost model demonstrated a strong predictive capability under both bootstrapped testing and blind testing conditions. In the CV training model the F1 scores ranged from Rating 2's score of 0.70 [CI 0.60 - 0.76] to Rating 1's score of 0.78 [CI 0.74 - 0.83] indicating a stable performance across the 3-class XGBoost model. This accuracy was similarly reflected in the bootstrapped testing F1 scores ranging from Rating 2's score of 0.77 [CI 0.67 - 0.87] to Rating 3's score of 0.84 [CI 0.75 - 0.93] (Table A4). Although the CIs ranged slightly wider in the testing results than in the training results, minimal variability and reasonable stability is still observed through this model.

When the 3-class XGBoost framework was applied to the blind test dataset the Recall scores turned out to be 0.90 for Class 1, and 0.42 for Class 2. The performance gap observed after applying blind testing highlights the realistic challenges in class-representation and distribution. However the blind test Recall score of 0.90 suggests that the model still maintains its predictive performance for the lower classes, consistent with the results seen in bootstrapped testing data. The recall score suggests that 90% of the individuals who were actually part of the lower sustainability class were correctly identified. Additionally, the Recall score of 0.42 for Class 2 further reflects this idea. Although this value is significantly lower than the 0.90 Class 1 Recall score, it indicates that 42% of the true instances were correctly identified in this class. Overall, these recall values show that the model performed well in accurately capturing lower sustainability class results in comparison to higher classes. This is a direct result of the minimized class-representation for the Class 3 of the model, as the support consisted of 4 individuals. Taking into account the lower representation within some of the classes, the overall strong accuracy ensures the generalizability of the XGBoost 3-class framework. The narrow confidence intervals seen in the CV training results further expresses the accuracy and reliability of the model's performance (Table A4). The discrepancy between blind testing and bootstrapped testing results is reflective of the survey-based dataset's limitations, and not a direct limitation of the model's stability.

The applied SHAP testing results further enhance the findings of the model by demonstrating the key input features that have the greatest influence on the model's output. Figure 3 reveals that sustainability-driven factors and consumption-based factors contain the highest absolute mean SHAP values, and therefore are ranked the highest among all the input features. As an example, EnvironmentalAwareness (CI 0.13) and MonthlyWaterConsumption (CI 0.11)—two of the highest ranked input features—demonstrate the most meaningful contributions as reflective of their mean SHAP values (Fig 3). The decline in SHAP values after these top indicators suggest that the model's decision making is concentrated on a small subset of features, while demographic variables contribute less to the overall sustainability outputs.

Similar to many prior studies, these findings align closely with other sustainability modeling research. As seen in the prior work of Wang et al. (2022), environmental awareness and resource-consumption behaviors are tied very closely with an individual's sustainability results. Additionally, other ensemble-based machine learning studies similar to ours highlight the optimized performance of gradient-boosted methods in comparison to baseline approaches to capture complex, non-linear behavioral patterns. This study, however, deviates from others due to its integration of SHAP-based methods to more accurately interpret the model's performance. Going beyond the general performance evaluation metrics, an explainable machine learning approach was taken to demonstrate that a smaller subset of data can produce the same optimized results with less complexity. The combination of a SHAP-based interpretability evaluation with feature importance analysis, the incorporation of an independent blind dataset to compare generalizability, and the application of both bootstrapping and cross validation techniques to assess model performance, collectively strengthen the findings of this study.

Limitations

Despite the strengths of this study in applying several machine-learning approaches, important limitations must be acknowledged. The study relies on self-reported, survey-based data, which is prone to bias due to varying participant interpretations of survey questions. Misreported behaviors and attitudes may create inconsistencies, reducing the reliability of input features and limiting the model's ability to learn generalizable patterns that improve predictive performance. Additionally, the blind testing dataset was relatively small and may not capture the complexity of a broader population. The testing dataset included 87 participants—around 20% of the original 500 individuals—which increases the probability of random variation and amplifies the risk of overfitting, where the model learns dataset-specific patterns that decrease generalizability. As a result, performance metrics may not reliably reflect how the model would perform on larger datasets, and outcomes may partly reflect sampling and class biases rather than solely predictive ability.

Another limitation involves the restricted number of input features used to train the model. Because the model was developed using a finite set of variables, it may not fully capture important behavioral and contextual predictors influencing sustainability outcomes. The exclusion of deeper behavioral factors may limit the model's ability to develop more complex relationships within the data, potentially reducing predictive performance and increasing variability in results.

The minimized demographic scope of participants—primarily from the United States, Canada, and India—also limits generalizability, as dominant patterns may reflect specific socioeconomic, cultural, or regional contexts rather than a broader global population. Furthermore, class imbalances within the datasets may have influenced evaluation metrics such as F1 scores, with underrepresented classes potentially skewing predictions toward majority classes. Although conversion to a 3-class model was implemented to mitigate this issue, imbalance remains a concern that may result in uneven performance. Lastly, the cross-sectional design of the study restricts the ability to observe changes in attitudes over time; incorporating longitudinal data, more diverse populations, expanded features, and improved class balance in future research would strengthen real-world applicability.

Conclusions

This study evaluates several machine learning models to predict an individual's sustainability score from behavioral data, leading to the development of an optimized model. By starting off with the evaluation of baseline models (Logistic Regression and KNN), and moving into the evaluation of balanced ensemble-based models (Random Forest, Extra Trees, LightGBM, and XGBoost), we utilized various optimization and validation methods to arrive at an optimal model choice. Cross-validation methods, bootstrapping approaches, confidence interval implementation, and reduced class representation were employed to ensure robust and reliable model evaluation.

Overall, the results indicate that the ensemble-based models, specifically XGBoost and LightGBM, consistently outperformed the baseline models. The comparison of individual F1 scores between the ensemble approaches seen in Tables A2 and A4, were used to assess performance across sustainability rating categories, rather than being dominated by majority-class predictions. Using these evaluation metrics, the 3-class XGBoost model was observed as the highest performing model due to its predictive performance and generalizability (Fig 2). Furthermore, the implementation of confidence interval analysis was utilized to reinforce the stability of the model by evaluating variation. The relatively narrow ranges observed across CV folds suggest consistent performance across various data samples.

Beyond performance improvements, a SHAP-based feature importance analysis approach was taken to meaningfully interpret the final model. The findings revealed that the following input features have the greatest impact on assigned sustainability scores: Transportation Mode, Environmental Awareness, Monthly Water Consumption, Monthly Electricity Consumption, Disposal Methods, and Home Size. Research indicates that a limited input subset of highly important features are sufficient to achieve higher predictive performances close to an optimal performance. The results underline the importance of using an explainable feature importance analysis to improve model efficiency while maintaining a strong predictive ability. Taken as a whole, the study demonstrates the implications of combining a balanced ensemble-based approach with explainable machine learning to produce generalizable predictions of sustainability behavior.

The findings of this study align with the prior research of Wang et al. (2022), as pro-environmental behaviors heavily correlate with the scale of environmental responsibility per individual. Additionally, like many other related works, this study demonstrates the effectiveness of ensemble-based machine learning models in behavioral and sustainability related predictive tasks. However, unlike many other studies that rely on large, unfiltered feature test sets, this study accounts for class imbalance by integrating a SHAP-based feature importance analysis directly into the model evaluation process. By demonstrating that a strong predictive performance can be achieved using a smaller subset of highly influential features, our study advances existing works that use explainability-driven feature selection processes.

Despite the overall strengths of the study, several setbacks and limitations should be acknowledged to adjust the focus of future works. The relatively small dataset size used for blind testing may limit the generalizability of the research findings, despite the use of CV, bootstrapping and CI analysis. This ties into the problem of class imbalances, which are present due to the lower class bias seen in classes 1 and 2. Therefore, the model demonstrated a strong predictive performance for the lower classes, reinforcing its effective predictive ability in this area. The reliance on self-reported survey data may promote response biases, affecting the accuracy of the predicted sustainability scores.

Works Cited

Todi, Shlok, and Patnala Akanksha. “A Study on Environmental Sustainability- Growth, Trends and Concerns.” IJFMR23022483, vol. 5, no. 2, 2023, ijfmr.com/papers/2023/2/2483.pdf.

Farhud, Dariush D. “Life Style and Sustainable Development.” Iranian Journal of Public Health, vol. 46, no. 1, 2017, p. 1, pmc.ncbi.nlm.nih.gov/articles/PMC5401917/.

Mansour, Fotouh R., and Alaa Bedair. “SUSTAIN as a Universal Scoring Tool for Assessing Sustainability Development Goals in African Energy Initiatives.” Scientific Reports, vol. 15, no. 1, 20 Nov. 2025, https://doi.org/10.1038/s41598-025-23521-x.

Wang, Qiaoling, et al. “Predictive Analysis of the Pro-Environmental Behaviour of College Students Using a Decision-Tree Model.” International Journal of Environmental Research and Public Health, vol. 19, no. 15, 31 July 2022, p. 9407, https://doi.org/10.3390/ijerph19159407.

Wang, Hairong. “Integrating Machine Learning into Life Cycle Assessment: Review and Future Outlook.” PLOS Climate, vol. 4, no. 10, 16 Oct. 2025, p. e0000732, https://doi.org/10.1371/journal.pclm.0000732.

Baehr, Julian, et al. “Predicting Product Life Cycle Environmental Impacts with Machine Learning: Uncertainties and Implications for Future Reporting Requirements.” Sustainable Production and Consumption, vol. 52, 9 Nov. 2024,sciencedirect.com/science/article/pii/S2352550924003178.

Sharma, Naveen. “Sustainable Lifestyle Rating Dataset.” Kaggle.com, 2024, kaggle.com/datasets/naveennas/sustainable-lifestyle-rating-dataset.

Sivakumar, Muthuramalingam, et al. “Trade-off between Training and Testing Ratio in Machine Learning for Medical Image Processing.” PeerJ Computer Science, vol. 10, 6 Sept. 2024, pp. e2245–e2245, peerj.com/articles/cs-2245/#, https://doi.org/10.7717/peerj-cs.2245.

Breiman, Leo. “Random Forests.” Machine Learning, vol. 45, no. 1, Oct. 2001, pp. 5–32.

Geurts, Pierre, et al. “Extremely Randomized Trees.” Machine Learning, vol. 63, no. 1, 2 Mar. 2006, orbi.uliege.be/bitstream/2268/9357/1/geurts-mlj-advance.pdf.

Chen, Tianqi, and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, vol. 1, no. 1, 13 Aug. 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

Ke, Guolin, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017.

Conciatori, Marco, et al. “Improving the Quality Evaluation Process of Machine Learning Algorithms Applied to Landslide Time Series Analysis.” Computers & Geosciences, vol. 184, Feb. 2024, p. 105531, https://doi.org/10.1016/j.cageo.2024.105531.

Arnold, Christian, et al. “The Role of Hyperparameters in Machine Learning Models and How to Tune Them.” Political Science Research and Methods, 5 Feb. 2024, cambridge.org/core/journals/political-science-research-and-methods/article/role-of-hyperparameters-in-machine-learning-models-and-how-to-tune-them/27296C04CF5935C55327F11BF4017371.

Kushan Sandunil, et al. “Effects of Tuning Decision Trees in Random Forest Regression on Predicting Porosity of a Hydrocarbon Reservoir. Case Study: Volve Oil Field, North Sea.” Energy Advances, 1 Jan. 2024, pubs.rsc.org/en/content/articlehtml/2024/ya/d4ya00313f, https://doi.org/10.1039/d4ya00313f.

Quadri, Mohammed Saim. “Random Forest: The Power of Many Trees.” Medium, 10 Sept. 2025, medium.com/@mohammedsaimquadri/random-forest-the-power-of-many-trees-4a619f3928b7.

Chen, Chao, and Andy Liaw. Using Random Forest to Learn Imbalanced Data. Jan. 2004.

Tayebi, Mohammed, and Said El Kafhali. “A Novel Approach Based on XGBoost Classifier and Bayesian Optimization for Credit Card Fraud Detection.” Cyber Security and Applications, Apr. 2025, p. 100093, https://doi.org/10.1016/j.csa.2025.100093.

Sibindi, Racheal, et al. “A Boosting Ensemble Learning Based Hybrid Light Gradient Boosting Machine and Extreme Gradient Boosting Model for Predicting House Prices.” Engineering Reports, 22 Nov. 2022, https://doi.org/10.1002/eng2.12599.

Gera, Divya. “Boosting Algorithms: AdaBoost, Gradient Boosting, XGB, Light GBM and CatBoost.” Medium, 9 Sept. 2020, medium.com/@divyagera2402/boosting-algorithms-adaboost-gradient-boosting-xgb-light-gbm-and-catboost-e7d2dbc4e4ca.

Zhang, Jesse, et al. Estimating Confidence Intervals on Accuracy in Classification in Machine Learning. 2019.

Tsamardinos, Ioannis, et al. “Bootstrapping the Out-of-Sample Predictions for Efficient and Accurate Cross-Validation.” Machine Learning, vol. 107, no. 12, 9 May 2018, https://doi.org/10.1007/s10994-018-5714-4.

Zhao, Chi, et al. “ShapG: New Feature Importance Method Based on the Shapley Value.” Engineering Applications of Artificial Intelligence, vol. 148, May 2025, p. 110409, https://doi.org/10.1016/j.engappai.2025.110409.

Wang, Huanjing, et al. “Feature Selection Strategies: A Comparative Analysis of SHAP-Value and Importance-Based Methods.” Journal of Big Data, vol. 11, no. 1, 26 Mar. 2024, https://doi.org/10.1186/s40537-024-00905-w.

Appendix

Table A1. Summary of train and test evaluation metrics using simple baseline models to predict 5 ratings of sustainability score

Model	Train/Test	Rating	Precision	Recall	F1	Samples
Logistic Regression	Train	1	0.82	0.82	0.82	77
		2	0.32	0.74	0.44	27
		3	0.69	0.63	0.66	81
		4	0.52	0.49	0.51	73
		5	0.78	0.64	0.70	141
	Test	1	0.89	0.80	0.84	20

		2	0.27	0.43	0.33	7
		3	0.58	0.75	0.65	20
		4	0.22	0.22	0.22	18
		5	0.78	0.60	0.68	35
KNN	Train	1	0.84	0.91	0.88	77
		2	0.46	0.41	0.43	27
		3	0.68	0.78	0.73	81
		4	0.64	0.68	0.66	73
		5	0.84	0.72	0.78	141
	Test	1	1.00	0.85	0.92	20
		2	0.25	0.14	0.18	7
		3	0.57	0.80	0.67	20
		4	0.33	0.28	0.30	18
		5	0.69	0.71	0.70	35

Table A2. Summary of train and test evaluation metrics with corresponding confidence intervals using tuned optimal models to predict 5 ratings of sustainability score

Model	Train/Test	Rating	Precision	Recall	F1	Samples
Random Forest *n_estimators = 1000* *min_samples_split = 10* *max_features = 1*	CV Train	1	0.87 CI [0.77 - 0.99]	0.78 CI [0.69 - 0.87]	0.82 CI [0.74 - 0.92]	77
		2	0.30 CI [0.00 - 0.95]	0.07 CI [0.00 - 0.20]	0.12 CI [0.00 - 0.33]	27
		3	0.69 CI [0.54 - 0.82]	0.59 CI [0.44 - 0.81]	0.63 CI [0.52 - 0.81]	81
		4	0.84 CI [0.53 - 1.00]	0.33 CI [0.10 - 0.46]	0.46 CI [0.16 - 0.63]	73
		5	0.60 CI [0.51 - 0.68]	0.97 CI [0.93 - 1.00]	0.74 CI [0.66 - 0.81]	141
	Bootstrapped Test	1	1.00 CI [1.00 - 1.00]	0.75 CI [0.55 - 0.95]	0.85 CI [0.71 - 0.97]	20
		2	0.66 CI [0.00 - 1.00]	0.14 CI [0.00 - 0.43]	0.23 CI [0.00 - 0.60]	7
		3	0.71 CI [0.59 - 0.85]	0.85 CI [0.70 - 1.00]	0.77 CI [0.67 - 0.88]	20
		4	0.99 CI [1.00 - 1.00]	0.22 CI [0.06 - 0.44]	0.35 CI [0.11 - 0.62]	18
		5	0.63 CI [0.56 - 0.70]	1.00 CI [1.00 - 1.00]	0.77 CI [0.71 - 0.82]	35
Xgboost *n_estimators = 100* *subsample = 0.6* *colsample_bynode = 0.6* *learning_rate = 0.1*	CV Train	1	0.89 CI [0.79 - 0.99]	0.81 CI [0.69 - 0.99]	0.84 CI [0.74 - 0.93]	77
		2	0.40 CI [0.21 - 0.56]	0.33 CI [0.17 - 0.64]	0.35 CI [0.20 - 0.60]	27
		3	0.70 CI [0.51 - 0.98]	0.64 CI [0.51 - 0.86]	0.65 CI [0.56 - 0.82]	81
		4	0.53 CI [0.43 - 0.75]	0.50 CI [0.24 - 0.66]	0.50 CI [0.31 - 0.60]	73
		5	0.71 CI [0.62 - 0.84]	0.77 CI [0.65 - 0.89]	0.73 CI [0.65 - 0.81]	141
	Bootstrapped Test	1	1.00 CI [1.00 - 1.00]	0.85 CI [0.70 - 1.00]	0.92 CI [0.82 - 1.00]	20
		2	0.91 CI [0.00 - 1.00]	0.29 CI [0.00 - 0.57]	0.42 CI [0.00 - 0.73]	7
		3	0.67 CI [0.54 - 0.82]	0.90 CI [0.75 - 1.00]	0.77 CI [0.65 - 0.87]	20

		4	0.85 CI [0.68 - 1.00]	0.61 CI [0.39 - 0.83]	0.71 CI [0.50 - 0.86]	18
		5	0.78 CI [0.69 - 0.89]	0.91 CI [0.83 - 1.00]	0.84 CI [0.77 - 0.92]	35
LightGBM *n_estimators = 100* *subsample = 0.5* *colsample_bytree = 0.6* *learning_rate = 0.3*	CV Train	1	0.82 CI [0.75 - 0.92]	0.82 CI [0.75 - 0.98]	0.82 CI [0.77 - 0.88]	77
		2	0.36 CI [0.17 - 0.65]	0.30 CI [0.17 - 0.57]	0.32 CI [0.12 - 0.54]	27
		3	0.69 CI [0.53 - 0.97]	0.63 CI [0.51 - 0.80]	0.64 CI [0.57 - 0.77]	81
		4	0.66 CI [0.51 - 0.97]	0.53 CI [0.30 - 0.73]	0.56 CI [0.38 - 0.66]	73
		5	0.70 CI [0.65 - 0.75]	0.78 CI [0.69 - 0.89]	0.74 CI [0.69 - 0.77]	141
	Bootstrapped Test	1	1.00 CI [1.00 - 1.00]	0.90 CI [0.75 - 1.00]	0.95 CI [0.86 - 1.00]	20
		2	0.46 CI [0.00 - 1.00]	0.15 CI [0.00 - 0.43]	0.21 CI [0.00 - 0.60]	7
		3	0.71 CI [0.58 - 0.86]	0.85 CI [0.65 - 1.00]	0.77 CI [0.64 - 0.89]	20
		4	0.72 CI [0.53 - 0.92]	0.56 CI [0.33 - 0.78]	0.62 CI [0.41 - 0.80]	18
		5	0.77 CI [0.67 - 0.87]	0.91 CI [0.80 - 1.00]	0.83 CI [0.75 - 0.91]	35
Extra Trees *n_estimators = 1000* *min_samples_split = 10* *max_features = 1*	CV Train	1	0.93 CI [0.77 - 1.00]	0.74 CI [0.63 - 0.86]	0.82 CI [0.77 - 0.92]	77
		2	0.47 CI [0.03 - 0.95]	0.15 CI [0.02 - 0.20]	0.22 CI [0.02 - 0.29]	27
		3	0.72 CI [0.53 - 0.80]	0.61 CI [0.50 - 0.86]	0.66 CI [0.51 - 0.82]	81
		4	0.86 CI [0.72 - 1.00]	0.30 CI [0.15 - 0.40]	0.43 CI [0.26 - 0.54]	73
		5	0.59 CI [0.52 - 0.65]	0.96 CI [0.90 - 1.00]	0.73 CI [0.66 - 0.79]	141
	Bootstrapped Test	1	1.00 CI [1.00 - 1.00]	0.80 CI [0.60 - 0.95]	0.89 CI [0.75 - 0.97]	20
		2	0.67 CI [0.00 - 1.00]	0.15 CI [0.00 - 0.43]	0.23 CI [0.00 - 0.60]	7
		3	0.73 CI [0.58 - 0.89]	0.80 CI [0.60 - 0.95]	0.76 CI [0.62 - 0.88]	20
		4	0.67 CI [0.25 - 1.00]	0.22 CI [0.06 - 0.39]	0.32 CI [0.09 - 0.54]	18
		5	0.62 CI [0.55 - 0.70]	0.97 CI [0.91 - 1.00]	0.76 CI [0.70 - 0.82]	35

Table A3. Summary of train and test evaluation metrics using simple baseline models to predict 3 ratings of sustainability score

Model	**Train/Test**	**Rating**	**Precision**	**Recall**	**F1**	**Samples**
Logistic Regression	Train	1	0.84	0.88	0.86	105
		2	0.79	0.78	0.79	153
		3	0.80	0.78	0.79	141
	Test	1	0.81	0.81	0.81	26
		2	0.68	0.69	0.68	39
		3	0.79	0.77	0.78	35
KNN	Train	1	0.82	0.84	0.83	105
		2	0.73	0.80	0.76	153
		3	0.82	0.72	0.77	141

	Test	1	0.71	0.65	0.68	26
		2	0.60	0.77	0.67	39
		3	0.81	0.60	0.69	35

Table A4. Summary of train and test evaluation metrics with corresponding confidence intervals using tuned optimal models to predict 3 ratings of sustainability score

Model	Train/Test	Rating	Preci-sion	Recall	F1	Samples
Random Forest *n_estimators = 500 min_samples_split = 20 max_features = 6*	CV Train	1	0.91 CI [0.85 - 0.99]	0.78 CI [0.72 - 0.81]	0.84 CI [0.79 - 0.87]	105
		2	0.79 CI [0.75 - 0.82]	0.63 CI [0.47 - 0.79]	0.69 CI [0.60 - 0.78]	153
		3	0.66 CI [0.58 - 0.74]	0.85 CI [0.79 - 0.89]	0.74 CI [0.70 - 0.80]	141
	Bootstrapped Test	1	0.95 CI [0.85 - 1.00]	0.74 CI [0.58 - 0.89]	0.83 CI [0.70 - 0.94]	26
		2	0.80 CI [0.69 - 0.92]	0.72 CI [0.56 - 0.85]	0.76 CI [0.65 - 0.86]	39
		3	0.69 CI [0.59 - 0.80]	0.88 CI [0.77 - 0.97]	0.78 CI [0.69 - 0.85]	35
Xgboost *n_estimators = 500 subsample = 0.8 colsample_bynode = 0.8 learning_rate = 0.3*	CV Train	1	0.84 CI [0.83 - 0.85]	0.73 CI [0.67 - 0.81]	0.78 CI [0.74 - 0.83]	105
		2	0.69 CI [0.66 - 0.71]	0.70 CI [0.55 - 0.80]	0.70 CI [0.60 - 0.76]	153
		3	0.73 CI [0.61 - 0.81]	0.77 CI [0.69 - 0.82]	0.75 CI [0.68 - 0.82]	141
	Bootstrapped Test	1	0.85 CI [0.71 - 0.96]	0.81 CI [0.65 - 0.96]	0.82 CI [0.71 - 0.93]	26
		2	0.77 CI [0.66 - 0.89]	0.77 CI [0.64 - 0.90]	0.77 CI [0.67 - 0.87]	39
		3	0.83 CI [0.72 - 0.94]	0.86 CI [0.74 - 0.97]	0.84 CI [0.75 - 0.93]	35
LightGBM *n_estimators = 500 subsample = 0.5 colsample_bytree = 0.8 learning_rate = 0.2*	CV Train	1	0.85 CI [0.80 - 0.93]	0.76 CI [0.76 - 0.76]	0.81 CI [0.78 - 0.84]	105
		2	0.70 CI [0.64 - 0.76]	0.75 CI [0.64 - 0.81]	0.72 CI [0.64 - 0.78]	153
		3	0.75 CI [0.68 - 0.84]	0.76 CI [0.72 - 0.82]	0.76 CI [0.70 - 0.81]	141
	Bootstrapped Test	1	0.80 CI [0.67 - 0.95]	0.77 CI [0.61 - 0.92]	0.78 CI [0.65 - 0.90]	26
		2	0.75 CI [0.64 - 0.88]	0.69 CI [0.54 - 0.82]	0.72 CI [0.60 - 0.83]	39
		3	0.77 CI [0.67 - 0.88]	0.86 CI [0.74 - 0.97]	0.81 CI [0.72 - 0.89]	35
Extra Trees *n_estimators = 500 min_samples_split = 20 max_features = 6*	CV Train	1	0.89 CI [0.85 - 0.93]	0.76 CI [0.68 - 0.81]	0.82 CI [0.78 - 0.85]	105
		2	0.75 CI [0.67 - 0.81]	0.66 CI [0.54 - 0.82]	0.70 CI [0.63 - 0.76]	153
		3	0.69 CI [0.59 - 0.82]	0.83 CI [0.69 - 0.90]	0.75 CI [0.71 - 0.80]	141
	Bootstrapped Test	1	0.89 CI [0.75 - 1.00]	0.62 CI [0.46 - 0.81]	0.72 CI [0.57 - 0.87]	26
		2	0.71 CI [0.60 - 0.81]	0.79 CI [0.67 - 0.92]	0.75 CI [0.65 - 0.84]	39
		3	0.72 CI [0.61 - 0.85]	0.77 CI [0.63 - 0.89]	0.74 CI [0.64 - 0.84]	35