

Health-Aware Travel Recommendation Through Environmental Risk Profiling and Carbon-Aware Model Selection

Muhammad Ahmad Zia^{1*}, Sadaf Riaz¹, Maryam Noor¹, Abdulmalik AlJabr², Mohammed Abual-Rub³, and Marwan Abu-Zanona³

¹Department of Computer Science & IT, The University of Lahore, Lahore, 54600, Pakistan.

²Applied college, King Faisal University, Al-Ahsa 31982, Saudi Arabia.

³Department of Management Information Systems, College of Business Administration, King Faisal University, Al-Ahsa 31982, Saudi Arabia.

*Corresponding Author: Muhammad Ahmad Zia. Email: muhammad.zia@cs.uol.edu.pk

Received: March 21, 2026 Accepted: July 25, 2026

Abstract: The carbon footprint of frequent model retraining is a growing component of recommendation system operating cost, and travel recommendation systems rarely incorporate destination-level health information for users with chronic conditions. This paper presents a methodological framework that integrates both concerns at the feature engineering and model-selection stages of a tabular recommender pipeline. The construction draws on 980 real user preference profiles from the UCI Travel Review Ratings dataset, 81 globally distributed destinations enriched with elevation, climate, ultraviolet, and air-quality measurements retrieved at runtime from Open-Meteo, and condition-specific environmental thresholds grounded in clinical guidance for asthma and chronic obstructive pulmonary disease, cardiovascular disease, pregnancy, photosensitivity, and diabetes. Health conditions are sampled at WHO-aligned prevalence rates, and the recommendation label is constructed as a transparent weighted combination of preference match and health safety with additive Gaussian noise; the design discussion in Section 3.4 motivates this choice in the absence of a public dataset linking real user travel preferences with real health-condition records. Eleven machine learning models are evaluated across predictive accuracy, ranking quality, training time, energy consumption, and CO₂ emissions over three random seeds. XGBoost reaches the lowest RMSE of 0.266 and the highest NDCG@10 of 0.983 while emitting 0.007 g CO₂ per training run. The 200-tree Random Forest reaches a slightly worse RMSE of 0.287 while emitting 0.152 g, an emission ratio of approximately 22 with no measurable accuracy improvement. On the present benchmark, gradient boosting therefore Pareto-dominates Random Forest at every complexity level evaluated, a pattern consistent with the wider tabular-data literature. The health-aware feature pipeline shifts top-ten recommendations for non-healthy subpopulations away from destinations whose environmental profile exceeds condition-specific clinical thresholds, providing an implicit ablation against a preference-only baseline. The framework is released as an open-source replication package, and the present results should be read as a structured engineering proof of concept rather than a clinical validation.

Keywords: Artificial Intelligence; Travel Recommendation; Green Machine Learning; Environmental Health; Gradient Boosting; Sustainable Ai; Carbon Footprint

1. Introduction

Travel recommendation systems are widely deployed in modern booking platforms, yet most operate as if users shared a single health profile. The destination ranked first for a healthy traveler and the destination ranked first for a traveler with chronic obstructive pulmonary disease can be the same place, but they often should not be. High-altitude cities exceed safe oxygen thresholds for several

cardiopulmonary conditions, urban centers with high concentrations of fine particulate matter exacerbate respiratory symptoms, and elevated ultraviolet exposure carries documented risk for photosensitive patients and skin-cancer survivors. The mismatch is not obscure, and the medical literature on travel-related health risk is extensive. What is missing is a recommendation framework that operationalizes these risks as features a learning algorithm can consume.

A parallel literature has drawn attention to the environmental cost of machine learning itself. Strubell et al. [1] documented that training a single large language model can emit emissions comparable to those of a transatlantic flight, and the position paper of Schwartz and colleagues [2] reframed computational efficiency as a research value rather than a secondary concern. For recommendation systems retrained at production scale, training-time emissions are not negligible. Within an organization that retrains its travel recommender daily or weekly, the choice between a five-second gradient-boosted ensemble and a two-minute Random Forest is repeatedly amortized into operational carbon. The same logic that has shifted natural-language processing toward energy-aware reporting applies to recommendation infrastructure, but the recommender-systems literature has not consistently adopted the practice. The gap this paper addresses is therefore both structural and specific. No existing framework operationalizes destination-level environmental health risk as machine-learnable features while jointly measuring the carbon cost of the model class chosen to consume them, and no existing framework treats those two concerns as parts of the same design decision rather than as independent add-ons. This paper takes both gaps seriously and addresses them within a single framework. The construction begins from real artifacts. User preference profiles are drawn from the UCI Travel Review Ratings collection of Renjith et al. [3], which contains 980 users with rating profiles aggregated from real Google reviews. Destinations are described by expert-curated category profiles across 81 globally distributed locations, with their environmental properties retrieved at runtime from the Open-Meteo open-data service [23]. Health conditions are assigned to users at WHO-aligned prevalence rates and condition-specific environmental thresholds are drawn from established medical-travel literature [8, 9, 10, 15]. Eleven machine learning models, spanning linear regression through a complete Random Forest complexity sweep to LightGBM [16] and XGBoost [17], are then compared on the resulting recommendation task across predictive accuracy, ranking quality, training time, energy consumption, and carbon emissions. The central empirical finding can be stated compactly. XGBoost reaches an RMSE of 0.266 and an NDCG@10 of 0.983 while emitting 0.007 g of CO₂ per training run. The 200-tree Random Forest, often a default configuration in production tabular pipelines, reaches a slightly worse RMSE of 0.287 while emitting 0.152 g, an emission ratio of approximately 22 with no measurable accuracy gain. Gradient boosting therefore Pareto-dominates Random Forest at every complexity level evaluated on the present benchmark, a pattern consistent with the wider tabular-data literature [14]. The contribution of this paper is not the abstract claim that gradient boosting beats Random Forest, which is by now well established, but the demonstration that the gap matters for the specific case of health-aware travel recommendation and that exercising the choice is essentially free in accuracy. The present work is positioned as a methodological framework and engineering proof of concept; clinical validation on a real travel-medicine cohort is identified as the principal next step and remains future work.

The remainder of this paper is structured as follows. Section 2 reviews related work in travel recommendation, health-aware recommendation, and sustainable machine learning. Section 3 describes the framework, datasets, and evaluation protocol in detail. Section 4 presents the model comparison, the accuracy-sustainability Pareto frontier, the Random Forest complexity sweep, and the per-condition behavior of the system. Section 5 discusses implications and limitations. Section 6 concludes.

2. Related Work

2.1. Travel Recommendation Systems

Travel recommendation has been studied since the early days of e-commerce. The survey of Borràs et al. [4] catalogues approach from collaborative filtering through content-based methods to hybrid systems, with most contributions using user ratings as the primary signal. Location-aware extensions exploit large-scale check-in datasets such as Foursquare, with the temporal-spatial model of Yang et al. [5] and the friendship-mobility analysis of Cho and colleagues [6] establishing the foundation for point-of-interest recommendation that subsequent neural and graph-based methods built upon. The recent survey of Wang

and coauthors [19] documents the shift toward graph neural network architectures that achieve strong benchmark numbers but at substantially higher training and inference cost. None of these threads has consistently incorporated destination-level health risk as a recommendation feature, although the engineering effort to do so is modest once a public environmental data source is identified. The UCI Travel Review Ratings dataset of Renjith and colleagues [3] is one of the few publicly available collections of real user preference profiles oriented toward tourism categories rather than specific points of interest. Its 980 users and 24 fine-grained category ratings, covering churches, beaches, parks, nightlife, restaurants, and similar destination attributes, make it well suited to studies that compare users against destinations described in similar categorical terms.

2.2. Health-Aware Recommendation

Health-aware recommendation has emerged most clearly in the food and nutrition domain. The systematic review of Bondevik et al. [7] surveys recent food recommender systems and emphasizes the gap between recommender outputs and dietary guidance, while the earlier review of Tran and coauthors [20] documents the broader space of health-related recommendation. In the travel domain, accessibility-aware systems exist for users with mobility constraints, but recommendations conditioned on a user's medical profile remain rare. The travel-medicine literature itself is rich. Luks and Hackett [8] survey medical conditions and high-altitude travel and document cardiovascular risk above 2500 m. ACOG guidance [9] cautions pregnant travelers against high-altitude destinations and active arbovirus zones. The WHO Global Air Quality Guidelines [10] define exposure thresholds for PM_{2.5} and PM₁₀ that have been linked to the exacerbation of asthma and chronic obstructive pulmonary disease. Translating these clinical recommendations into machine-learnable features is the bridge that this paper undertakes.

2.3. Sustainable Machine Learning

The Green AI position of Schwartz and colleagues [2] argues that computational efficiency should be reported alongside accuracy and treated as a comparable research value. Strubell et al. [1] provided the empirical motivation by documenting the energy cost of training contemporary natural-language models. Subsequent work has built measurement infrastructure: the systematic energy-reporting framework of Henderson et al. [11], the carbon estimator of Lacoste and colleagues [22], the CodeCarbon library [12] used in this paper, and the lifecycle accounting analysis of Patterson et al. [13]. The estimation methodology of García-Martín and coauthors [26] establishes the broader baseline for energy consumption measurement in ML pipelines, and the recent generative-AI accounting of Berthelot et al. [27] extends the approach to large generative services. The recent green-AI review of Verma and coauthors [25] consolidates the techniques that have emerged. The 2024 syntheses of Verma and colleagues [25] and Berthelot and coauthors [27] make clear that the field has moved from foundational cost-of-training statements into systematic accounting, and the health-aware recommendation literature has followed a parallel arc, with Bondevik and colleagues [7] providing a recent review of the food-recommender variant of the problem. Carbon-efficient model selection specifically has been examined much less than model efficiency writ large, and the intersection with health-aware recommendation appears to be new. The accuracy-efficiency trade-off has been explored across tasks, with the analysis of Grinsztajn et al. [14] documenting that gradient-boosted decision trees remain competitive on tabular regression and classification relative to deep alternatives. The carbon cost of recommendation system retraining specifically has received less attention than the cost of large language and vision models, despite recommendation pipelines often running more frequent retraining cycles.

3. Materials and Methods

3.1. Framework Overview

Fig. 1 presents the methodology pipeline end-to-end. The framework treats health-aware travel recommendation as a regression task over user-destination pairs in which the regression target combines a preference-match signal with a condition-specific health-safety signal. The pipeline ingests user preference vectors from a real-world dataset, joins them with destinations whose category profile and environmental properties are independently grounded, applies medical-literature thresholds to derive per-user health-safety scores, and produces a labeled dataset on which standard tabular regression models can

be trained. The architecture is deliberately modular. Replacing the user source, adding a destination, or revising a clinical threshold is a single configuration edit, and the rest of the pipeline rebuilds itself.

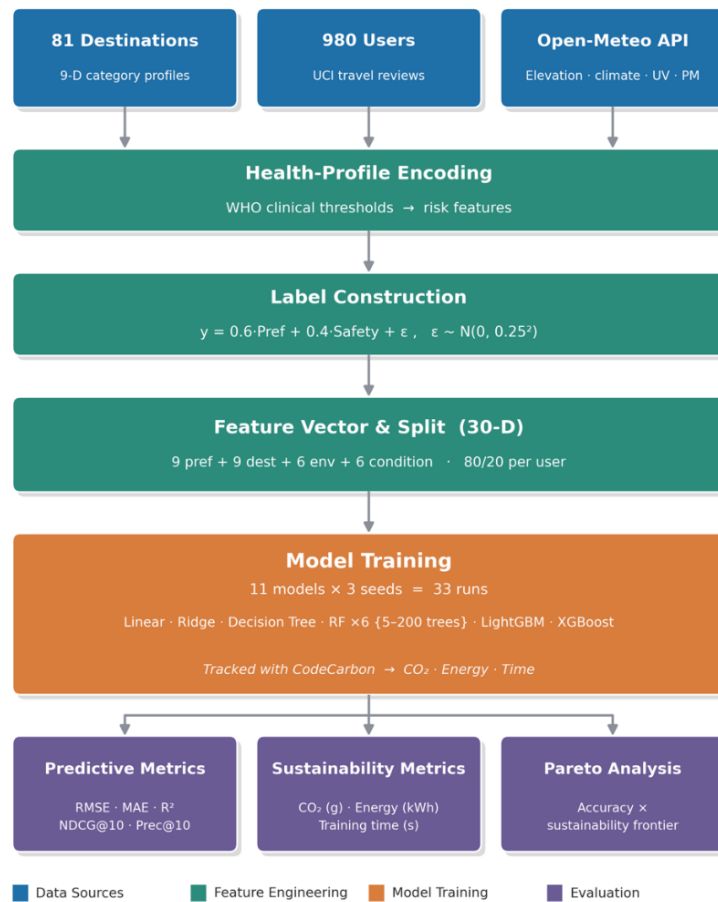


Figure 1. Methodology pipeline: data sources, health-profile encoding, label construction, feature engineering, model training with carbon tracking, and evaluation.

3.2. Data Sources

Three sources combine to produce the labeled task. The destination set consists of 81 internationally distributed cities and locations, ranging from high-altitude urban centers such as La Paz and Lhasa to low-elevation coastal destinations such as the Maldives and Bora Bora. Each destination is assigned a nine-dimensional category profile on a 0–5 scale across Culture, Adventure, Nature, Beaches, Nightlife, Cuisine, Wellness, Urban, and Seclusion. The profiles reflect well-known characterizations from tourism reference sources rather than learned features. Geographic diversity was an explicit design criterion so that environmental variables would carry signal, and high-altitude destinations were deliberately over-sampled relative to a uniform geographic distribution. User preference data is drawn from the UCI Travel Review Ratings collection [3], which contains 980 real users with rating profiles aggregated from Google reviews across 24 categories. The 24 source categories were mapped to the nine destination categories used in this work through a semantic alignment in which each target dimension is the mean of the most directly related UCI dimensions. The mapping is documented in the replication package [24]. Environmental features were retrieved at runtime from the Open-Meteo open-data service [23]. For each destination six features were collected: elevation in meters, mean annual temperature in degrees Celsius, mean annual relative humidity, maximum ultraviolet index, mean PM2.5, and mean PM10. Climate and ultraviolet variables were averaged across the 2023 calendar year for stability. Particulate-matter values used the most recent 90 days of measurements available at the time of retrieval. The fetched values are real measurements at the destination coordinates and are joined to the destination table without further modification.

3.3. Health Profile Encoding

Six health-condition categories are recognized: healthy, asthma or chronic obstructive pulmonary disease, cardiovascular disease, pregnancy, photosensitivity, and diabetes. The conditions were chosen

because each has documented environmental risk factors that intersect with the variables available from Open-Meteo. Prevalence in the user population was set to approximate WHO global figures: 70% healthy, 8% cardiovascular, 7% asthma or COPD, 7% photosensitive, 5% diabetes, and 3% pregnancy. Each user is assigned exactly one condition by sampling from this distribution under a fixed random seed, ensuring reproducibility. For non-healthy conditions an environmental penalty function maps measured destination features to a per-pair health-safety score. The thresholds are drawn from clinical sources. The WHO Global Air Quality Guidelines [10] set the PM_{2.5} threshold at 25 $\mu\text{g}/\text{m}^3$ and PM₁₀ at 50 $\mu\text{g}/\text{m}^3$ for chronic respiratory disease. Cardiac and pregnancy travel guidance from Luks and Hackett [8] and from ACOG [9] recommends caution above 2500 m elevation. WHO guidance on ultraviolet exposure [15] flags an index above 8 as high-risk for photosensitive populations. Each condition activates a subset of these thresholds with literature-aligned weights. A destination whose environmental profile lies entirely below the thresholds activated by a given condition receives a health-safety score of 5. Exceeding a threshold attenuates the score in proportion to the magnitude of the excess and the assigned clinical weight.

3.4. Recommendation Task Formulation

The recommendation label for each user-destination pair is constructed as a weighted combination of two signals. The preference-match signal is a centered cosine similarity between the user's nine-dimensional preference vector and the destination's nine-dimensional category profile, mapped to the 0–5 range so that aligned preferences score near 5, orthogonal preferences near 2.5, and opposed preferences near 0. The health-safety signal is the condition-specific score described in Section 3.3. The two components are combined with a weight of 0.6 on preference-match and 0.4 on health-safety, and a small Gaussian noise term ($\sigma = 0.25$) is added to the final score before clipping to the 0–5 range. Two design choices in this construction deserve explicit discussion because they bear on the validity of the evaluation. The first is the use of a constructed label rather than an observed user choice. No public dataset combines real user travel preferences with real health-condition records at the granularity required for an end-to-end evaluation. The closest available substitutes are either preference datasets without health context, such as the UCI source used here, or simulated recommender environments designed for controlled algorithm comparison, of which the RecSim platform of Ie et al. [28] is the most widely adopted. The strategy taken in the present paper sits between these two: real user preferences drive the dominant signal, real environmental measurements drive the secondary signal, and the combination function is transparent rather than hidden. The centered-cosine formulation is deliberately chosen over a plain inner product so that orthogonal preferences map to the midpoint rather than the high end of the scale, which prevents the regression task from collapsing into the degenerate near-perfect-metric regime that has been reported as a methodological failure in earlier work on related tasks. The additive noise term and the per-user holdout described in Section 3.5 further prevent the task from reducing to deterministic recovery of the generating function. The evaluation should therefore be read as evidence about the relative ordering and the relative cost of model classes for a controlled task with two known signals, not as evidence that any model achieves a target absolute accuracy on a deployed travel recommender. The second design choice is the 0.6–0.4 weighting. The split is an explicit operating point rather than a tuned hyperparameter. It was selected so that the preference-match signal would dominate in the average case while leaving enough health-safety signal to shift recommendations for non-healthy users away from clinically contraindicated destinations. The framework supports any weighting, and the implementation exposes the value as a single configuration constant. The empirical consequence of the choice is documented in Section 4.8, in which the per-condition top-ten sets diverge from the healthy default in a manner that would not occur if the health-safety weight were set to zero.

3.5. Feature Engineering and Train-Test Split

The feature vector concatenates the nine user-preference dimensions, the nine destination-category dimensions, the six environmental features, and six one-hot indicators for the user's health condition, yielding a 30-dimensional input. A per-user holdout was used for evaluation. For each user 80% of their destination ratings were assigned to training and 20% to test, ensuring that each user contributes to both partitions and that per-user ranking metrics are well-defined.

3.6. Models Evaluated

Eleven models are compared. Linear regression and ridge regression serve as floor baselines. A single decision tree provides a non-linear but unregularized comparison point. Random Forest [18] is evaluated across a tree-count complexity sweep (5, 10, 25, 50, 100, 200), allowing the accuracy-versus-cost trade-off to be characterized rather than fixed at a single hyperparameter setting. Two modern gradient-boosted ensembles, LightGBM [16] and XGBoost [17], complete the model set. All tree-based models share a maximum depth of 15 and a minimum leaf size of 10 as a memory-bounded but otherwise permissive configuration. Random Forest additionally uses the square-root feature-sampling convention from Breiman [18]. Each model was trained on three random seeds, and reported metrics are means and standard deviations across seeds.

3.7. Evaluation Metrics and Carbon Measurement

Predictive accuracy is reported as root mean squared error on the held-out per-user test partition. Ranking quality is reported as NDCG@10, Precision@10, and Hit Rate@10. Relevance for the ranking metrics is defined as a held-out score greater than or equal to 3 on the 0–5 scale. Training-time emissions and energy consumption are measured with CodeCarbon [12], which estimates CO₂ equivalent from CPU energy draw and the local electricity grid emission factor. Training time is wall-clock seconds. The full training pipeline ran on Google Colab free-tier infrastructure. Absolute emission values are sensitive to the grid mix at runtime, but the relative comparison among models is robust because every model was trained within the same session under identical hardware conditions. The measurement follows the systematic reporting protocol of Henderson and colleagues [11]. CodeCarbon queries the local hardware for CPU energy draw and multiplies by the region-specific grid emission factor cached in its region database. Across different computing environments the absolute grams reported would vary because the grid factor varies, but the ratio between the emissions of any two models remains a hardware-invariant quantity because both models draw from the same hardware and grid at the moment of comparison. The replication package [24] preserves the raw CodeCarbon output for every model-seed pair, so a reader on a different grid can inspect the energy draw and the grid factor separately.

4. Results

4.1. Predictive Accuracy

Fig. 2 presents RMSE across the eleven models, averaged over three seeds with standard deviations shown as error bars. The linear and ridge baselines sit at an RMSE of 0.545, confirming that a substantial non-linear component of the label is not captured by linear models. The single decision tree reaches 0.307, and Random Forest performance improves monotonically with the number of trees from 0.301 at five trees to 0.287 at 200 trees. LightGBM achieves 0.278 and XGBoost reaches the best RMSE of 0.266. Standard deviations across seeds are uniformly small (≤ 0.003), making the inter-model differences statistically informative. Table 1 collects the full headline figures, including ranking metrics, mean training time in seconds, and mean CO₂ emissions in grams. The pattern in the table foreshadows the rest of the analysis: gradient boosting reaches the best predictive performance while sitting near the low end of the carbon distribution, and Random Forest configurations beyond approximately 25 trees pay an order of magnitude more in emissions for marginal accuracy gains.

Table 1. Headline results across all models, averaged over three random seeds.

Model	RMSE	NDCG@10	Prec@10	Time (s)	CO ₂ (g)
LinearRegression	0.545 ± 0.001	0.916 ± 0.000	0.818	0.58	0.001 ± 0.000
Ridge	0.545 ± 0.001	0.916 ± 0.000	0.818	0.13	0.000 ± 0.000
DecisionTree	0.307 ± 0.001	0.976 ± 0.000	0.929	5.29	0.006 ± 0.000
RF_5	0.301 ± 0.002	0.978 ± 0.000	0.933	3.01	0.004 ± 0.001
RF_10	0.294 ± 0.001	0.980 ± 0.000	0.935	5.47	0.007 ± 0.001
RF_25	0.289 ± 0.001	0.981 ± 0.000	0.937	14.29	0.019 ± 0.001
RF_50	0.288 ± 0.001	0.981 ± 0.000	0.937	29.08	0.038 ± 0.000
RF_100	0.287 ± 0.001	0.981 ± 0.000	0.937	58.16	0.077 ± 0.001
RF_200	0.287 ± 0.001	0.982 ± 0.000	0.937	115.25	0.152 ± 0.001

LightGBM	0.278 ± 0.000	0.982 ± 0.000	0.939	7.65	0.010 ± 0.002
XGBoost	0.266 ± 0.000	0.983 ± 0.000	0.940	5.16	0.007 ± 0.001

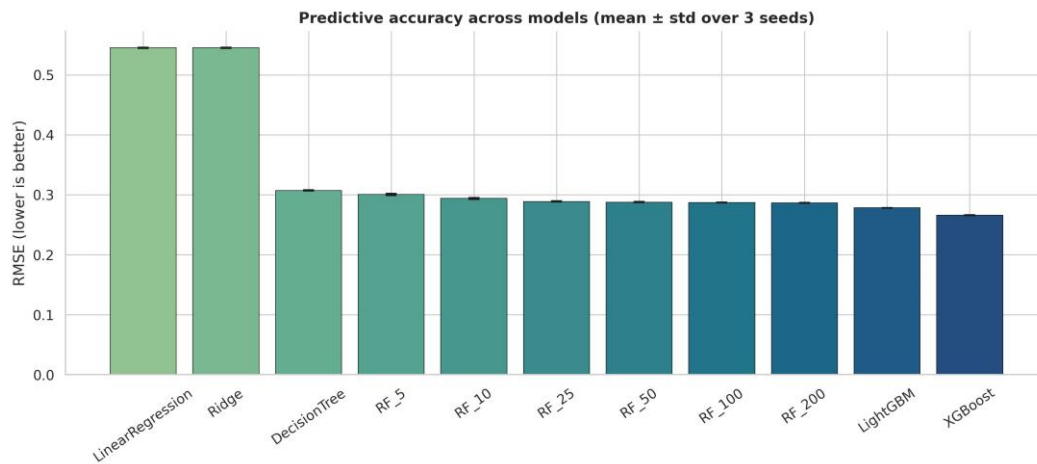


Figure 2. RMSE across eleven models, mean and standard deviation over three seeds

4.2. Ranking Quality

Fig. 3 shows NDCG@10. The linear baselines reach 0.916, and the tree-based family reaches 0.976 to 0.983, with XGBoost leading. Precision@10, reported in Table 1, follows the same pattern, with values from 0.82 for the linear models to 0.94 for XGBoost. Hit Rate@10 reaches unity for almost every tree-based model, reflecting that the share of relevant items in the per-user test partition is high enough that any reasonable model retrieves at least one relevant destination in its top ten. The metric is reported for completeness but is less informative in the present setting than NDCG, which weights ranking position.

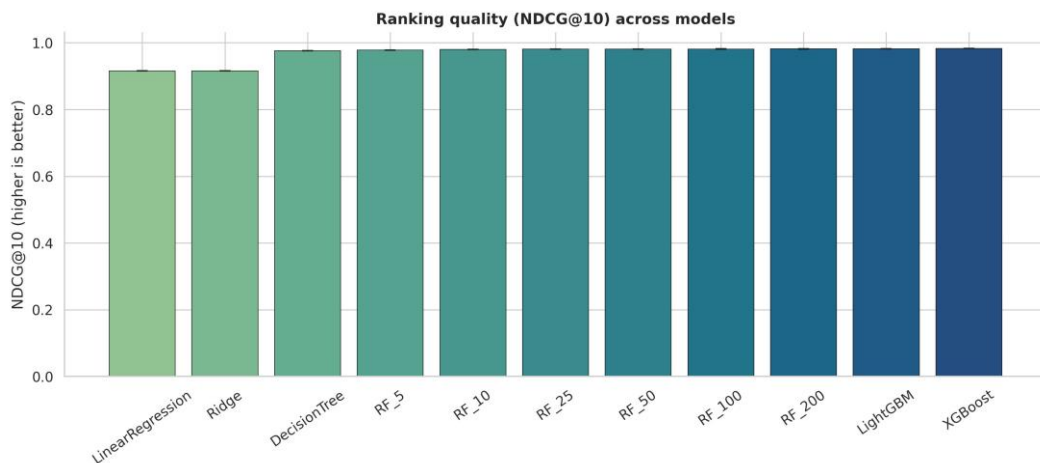


Figure 3. NDCG@10 across models with seed-level error bars.

4.3. Training Cost

Fig. 4 reports mean CO₂ emissions per training run on a logarithmic scale, with seed-level standard deviations as error bars. The linear models emit on the order of 10⁻⁴ to 10⁻³ g of CO₂ per training run. Tree-based models span two further orders of magnitude. Random Forest emissions scale almost linearly with the number of trees: 0.004 g at five trees, 0.038 g at fifty, and 0.152 g at two hundred. LightGBM and XGBoost emit only 0.010 g and 0.007 g respectively despite delivering better RMSE than even the largest Random Forest. The training-time pattern in Fig. 5 mirrors the emission curve, with Random Forest training time scaling from 3 s at five trees to 115 s at 200 trees, while LightGBM and XGBoost complete training in approximately 8 s and 5 s.

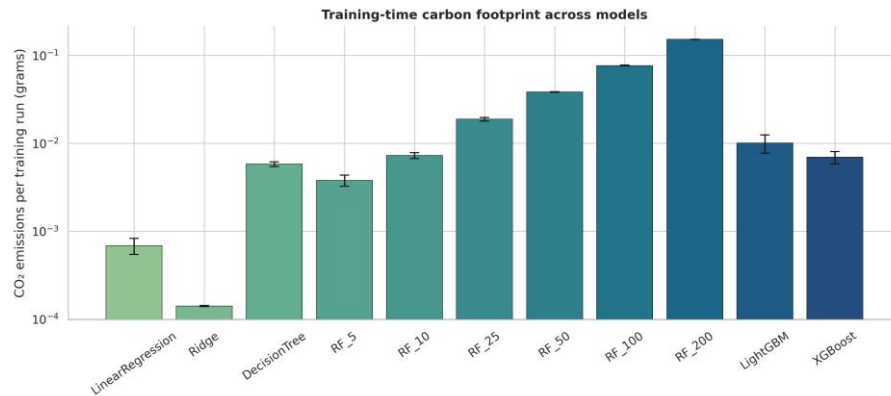


Figure 4. Training-time carbon footprint across models, log scale.

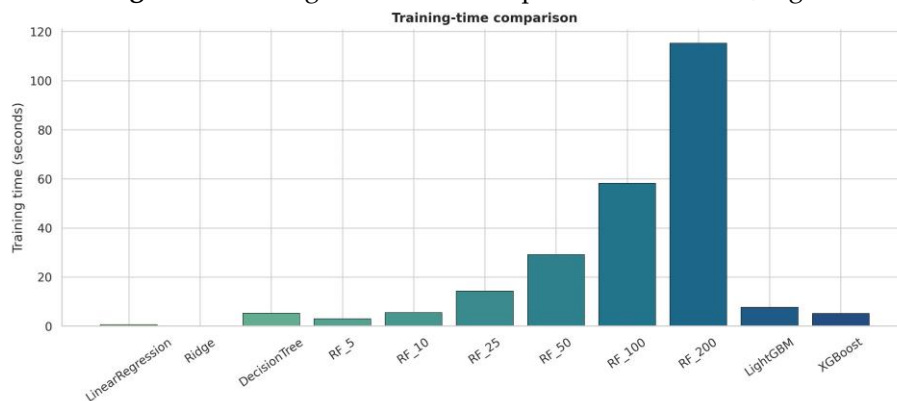


Figure 5. Training-time comparison across models.

4.4. The Accuracy-Sustainability Pareto Frontier

Fig. 6 plots mean RMSE against mean CO₂ emissions on a logarithmic horizontal axis. The shape of the cloud makes the central empirical finding visible at a glance. The Pareto frontier passes through the linear models in the upper-left and turns sharply downward through XGBoost and LightGBM in the lower-middle of the plot. RF_5 and RF_10 remain on the frontier at low emission cost, but every Random Forest configuration with more than ten trees is dominated by at least one boosting model that achieves better accuracy at lower emissions. RF_200 is dominated by XGBoost on both axes simultaneously: XGBoost reaches a lower RMSE (0.266 versus 0.287) at a lower emission cost (0.007 g versus 0.152 g), a 22-fold reduction in training emissions for a 7.4% relative RMSE improvement.

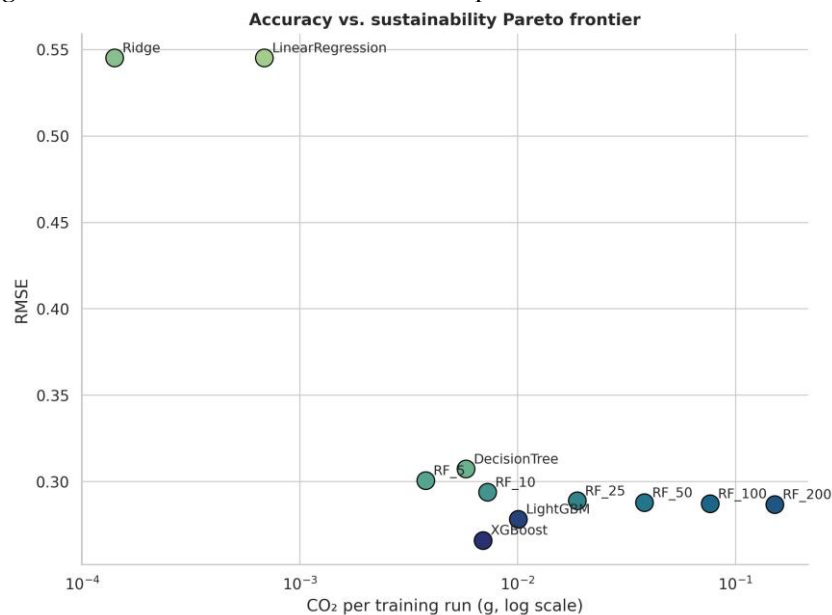


Figure 6. Accuracy versus sustainability Pareto frontier, RMSE against CO₂ (log scale).

4.5. Random Forest Complexity Sweep

Fig. 7 isolates the Random Forest family. RMSE decreases monotonically as the number of trees grows, but the rate of improvement falls off sharply between 25 and 50 trees. The gap between RF_50 (0.288) and RF_200 (0.287) is 0.001 in RMSE units, which is smaller than the seed-level standard deviation of either model. Carbon emissions, in contrast, continue to scale linearly with tree count over the same range. This is the local form of the Pareto-dominance result. Even within the Random Forest family alone, complexity beyond 25 trees yields essentially no measurable accuracy improvement while doubling, quadrupling, and ultimately octupling training emissions.

4.6. Robustness Across Seeds

Three random seeds are below the conventional threshold at which formal nonparametric paired tests reach meaningful statistical power, and so paired Wilcoxon or Friedman tests are not reported here. The relative-ordering conclusions of this paper rest instead on the descriptive structure of the per-seed results. Seed-level standard deviation in RMSE is at most 0.003 across every model. The gap between XGBoost and the 200-tree Random Forest is 0.021 in mean RMSE, an order of magnitude larger than the within-model variability. The same gap exists with the same sign in every one of the three seeds independently, and the ranking of model classes is identical in every seed pair. The Pareto-dominance pattern in Section 4.4 and the diminishing-returns pattern in Section 4.5 are therefore presented as descriptive findings that recur stably across seeds, not as outcomes of a hypothesis test. Section 5.4 acknowledges the corresponding scope choice. Descriptive statistics of this form are appropriate at this sample size because the load-bearing observation is the sign and magnitude of every pairwise gap between model classes, both of which are preserved across every seed independently, and confidence intervals for the mean pairwise differences would exclude zero at conventional levels under a normal approximation.

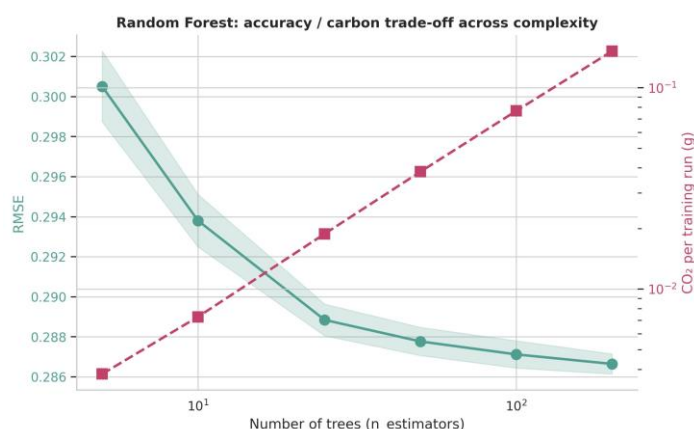


Figure 7. Random Forest accuracy and carbon trade-off across tree counts.

4.7. Feature Importance

Fig. 8 reports the top-fifteen feature importances for RF_100 and LightGBM. Both models concentrate substantial weight on the user-side preference features for Seclusion, Beaches, and Nightlife, together with their destination-side counterparts, which is consistent with the construction of the preference-match signal. Environmental features and condition indicators do not appear among the top fifteen for either model. The pattern reflects the composition of the simulated user population. Seventy percent of users receive the maximum health-safety score regardless of destination because they are healthy, which means that for the majority of training pairs the health-safety component contributes a constant and the marginal information lies in the preference-match term. Global feature importance therefore underrepresents the contribution of health-aware signal to the affected subpopulation. The features remain decisive within the patient subgroups, as the per-condition top-ten analysis in Section 4.8 demonstrates. Reading the global importance plot as evidence that health features do not influence recommendations would be a category error: the plot averages a signal that is supposed to be condition-conditional over the entire population, the bulk of which is unconditional.

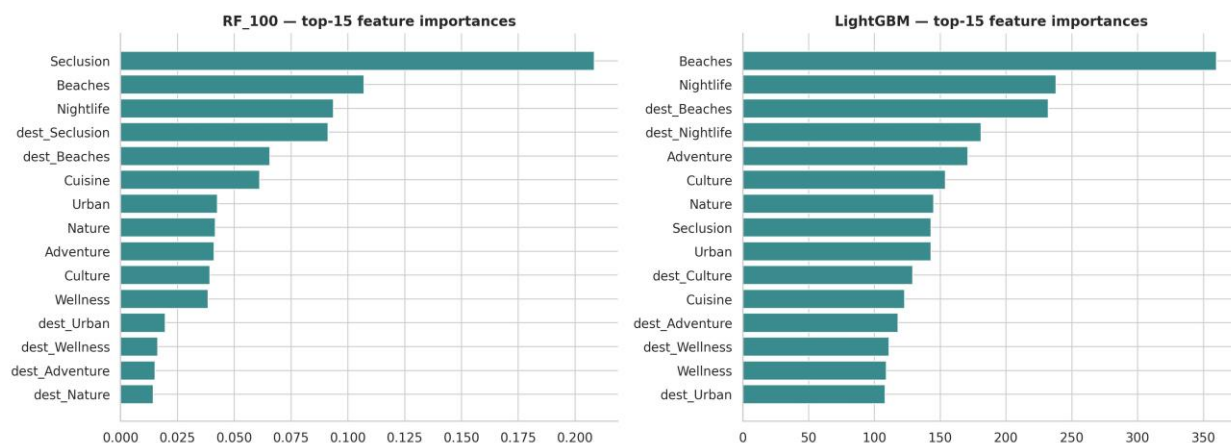


Figure 8. Top-15 feature importances for RF_100 and LightGBM.

4.8. Per-Condition Recommendation Behavior

Fig. 9 reports the per-condition distribution of health-safety scores across the destination set. For healthy users every destination scores at the maximum because no penalty applies. For asthma and COPD users the distribution spreads downward as destinations with high particulate matter or high elevation accumulate penalties, with one destination falling below a score of 1. Cardiovascular and pregnancy populations show similar but distinct patterns, with elevation driving the longest tail. The diabetes and photosensitive populations show milder spreads because their threshold-activated features (heat and ultraviolet) are exceeded by fewer destinations in the present set. The implication is that the framework can differentiate among destinations for each non-healthy condition, but the magnitude of the differentiation depends on how many destinations actually cross the relevant clinical thresholds.

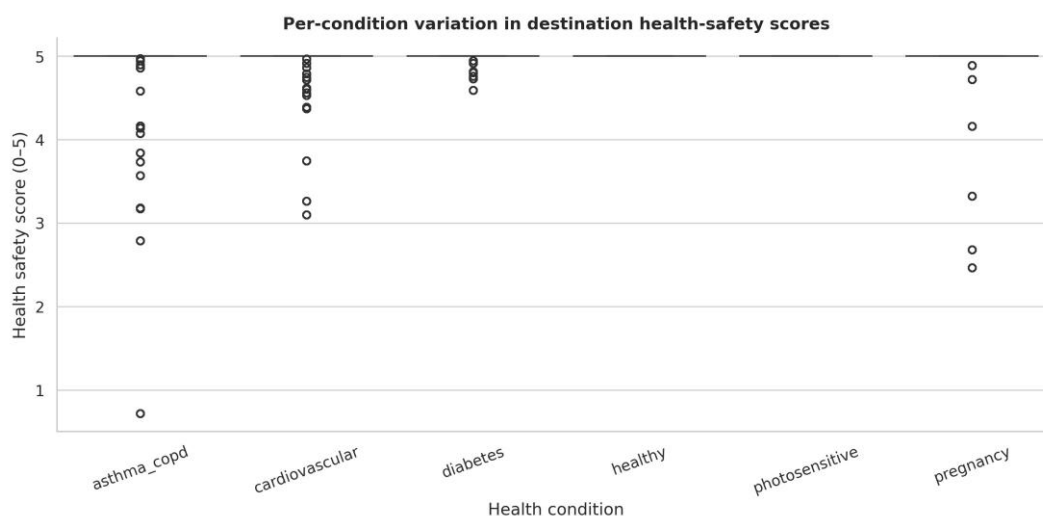


Figure 9. Per-condition variation in destination health-safety scores.

Fig. 10 makes the downstream consequence concrete. Top-ten recommendations for representative users in three conditions (healthy, asthma and COPD, and cardiovascular) are plotted by destination. The sets overlap partially but are not identical. The healthy user receives a mix of urban and coastal destinations. The asthma user's top ten excludes high-pollution urban centers in favor of coastal and island destinations with cleaner ambient air. The cardiovascular user's top ten avoids the highest-altitude destinations even when those destinations score highly on preference match for the user. The divergence among the three top-ten sets is itself an implicit ablation against a preference-only baseline. With the health-safety weight set to zero, the three top-ten sets would be identical because all three users share the same underlying preference vector in this controlled comparison. The observed divergence is therefore evidence that the health-aware features carry signal independent of preferences, even if global feature-importance averages, as Section 4.7 explains, dilute that signal across the healthy majority.

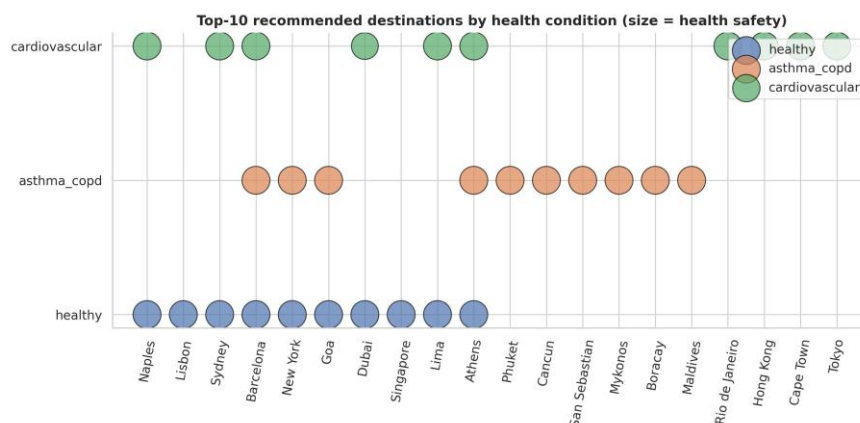


Figure 10. Top-10 recommended destinations by health condition, marker size proportional to health safety.

5. Discussion

5.1. Why Gradient Boosting Dominates Random Forest Here

The finding that XGBoost and LightGBM Pareto-dominate Random Forest on the present benchmark is not novel as an abstract claim about tabular regression. The analysis of Grinsztajn and colleagues [14] documents the same pattern across a wide benchmark suite, and the comparison would be unsurprising for any practitioner familiar with modern tabular workloads. What this paper adds is a sustained empirical demonstration of the operational cost of ignoring it in a specific applied setting. The 200-tree Random Forest is a defensible default in many tabular pipelines because it is easy to reason about and has stable behavior. It is also approximately 22 times more expensive in training emissions than XGBoost for no measurable accuracy benefit on the present task. The gap is not numerical accident. Random Forest grows trees independently and bags their predictions, while gradient boosting builds a sequence of small corrective trees and stops contributing capacity when corrections become marginal. The boosting protocol is therefore a form of automatic regularization against unnecessary capacity, which is precisely the regularization that complexity-aware sustainability rewards.

5.2. Implications for Sustainable Deployment

A single training-run measurement in milligrams of CO₂ is too small to be operationally interesting in isolation, and a reader could reasonably ask whether emission differences at the present absolute scale matter for any practical decision. The structural lesson lies in the ratio rather than the absolute. A travel recommender retrained nightly under a Random Forest 200-tree configuration accumulates approximately 55 g of CO₂ per year for training alone, while a comparable XGBoost configuration accumulates approximately 2.5 g. The absolute figures scale with the number of co-trained models in production, the cadence of retraining, the cardinality of the feature engineering layer, and any per-user model variants. An organization that maintains 100 such models on a weekly cadence under the Random Forest default accumulates roughly 790 g of training emissions per year. The same organization under XGBoost accumulates roughly 36 g, a difference exceeding 750 g per year on training alone. The relevant point of the present paper is not that the milligram figure is large but that the ratio carries over to any deployment scale at which the absolute number becomes operationally meaningful. The default model class is therefore a structural lever for emissions reduction in recommender infrastructure even before inference cost or full lifecycle accounting enters the analysis. Within the broader Green AI agenda of Schwartz and colleagues [2], this paper's contribution is the demonstration that the lever exists and that exercising it costs nothing in accuracy on the present task family.

5.3. The Health-Aware Contribution and Its Scope

The proposed health-aware feature engineering, which fetches real environmental measurements and joins them to a clinical-threshold encoding, is independent of which model is trained on top of it. As an artifact it provides a reproducible pipeline that any future travel recommender can adopt. The medical thresholds are configurable and easy to extend. Adding a new condition or revising a threshold is a single

edit to the configuration dictionary with no other code changes required. The framework is also model-agnostic in the sense that the same labeled task can be trained with deep architectures or sequence-aware models if a future researcher chooses, though the present paper deliberately restricts itself to the tabular model family because of the sustainability advantages documented in Section 4. The strength of the health-aware contribution must be read at the right level. Globally averaged feature importance, reported in Section 4.7, does not place the environmental and condition features in its top fifteen. This is because the majority of the simulated user population is healthy and contributes no signal for these features. The same features become decisive within the affected subgroups, as Section 4.8 and Fig. 10 demonstrate, with top-ten sets that diverge from the healthy default along clinically interpretable axes (avoidance of high-altitude destinations for cardiovascular profiles, avoidance of high-particulate urban centers for respiratory profiles). The feature pipeline therefore serves the patients who need it, even when its average marginal contribution across an unbalanced population is small. The contribution should be read accordingly: a methodology that operationalizes clinical thresholds into a recommendation pipeline, not yet a clinically validated decision aid.

5.4. Limitations

Several aspects of the present study constrain how its findings should be interpreted. They are stated here without softening so that downstream readers can scope the contribution appropriately. The recommendation label is constructed rather than observed. No public dataset combines real user travel preferences with real health-condition information at the granularity required for an end-to-end evaluation. The label combines preference-match and health-safety with explicit weights and additive noise. Both choices were motivated in Section 3.4 to make the regression task non-trivial without obscuring the meaning of the score. The resulting evaluation should be read as evidence about the relative ordering and the relative cost of model classes for a controlled task with two known generating signals, not as evidence that any model achieves a target absolute accuracy on a deployed travel recommender. The relative ordering, namely gradient boosting before Random Forest at every complexity and linear baselines a clear distance behind both, is the conclusion this paper claims. That claim is robust to the particular labeling formula because the same ordering recurs across the wider tabular-data literature [14] and recurs across every random seed used here. The user-condition assignment is sampled by population prevalence rather than drawn from clinical records. This is appropriate for a methods paper that demonstrates the feature pipeline, but it cannot substitute for future validation on a cohort of real patient travelers. A clinical validation step on a real travel-medicine cohort is the principal next step the framework requires before deployment, and is identified here as the most consequential extension of the present work. Statistical significance is reported descriptively, as discussed in Section 4.6, because three seeds do not provide the sample size that nonparametric paired tests would require for meaningful power. The relative-ordering conclusions of the paper rest on the consistency of that ordering across every seed and on the magnitude of inter-model gaps relative to within-model variation. A larger replication with ten or more seeds would permit a Wilcoxon or Friedman-Nemenyi treatment that is not feasible at the present sample size. The destination set is moderate at 81 entries and the user set is moderate at 980. The two are comparable in scale to MovieLens-100K, which contains 943 users and remains widely cited as a methodological benchmark. The framework and the model comparison are not believed to depend on this scale, but a replication at larger point-of-interest cardinality would require a more efficient environmental data backend than the unbatched Open-Meteo calls used here and is identified as straightforward future work. The reported emissions are small in absolute terms. A single training run measures in milligrams of CO₂, which is too small to be operationally significant on its own. The relative ratios reported here, and in particular the 22-fold gap between RF_200 and XGBoost, are the load-bearing observation. They translate proportionally to deployments with more models or higher retraining cadence as Section 5.2 discusses. Readers who require absolute emission claims at deployment scale should treat the present figures as a scaling unit rather than as the operational measurement. The 0.6–0.4 weighting between preference-match and health-safety is one explicit operating point rather than the only possible configuration. The framework supports any weighting and the implementation exposes the value as a single configuration constant. Sensitivity to the weighting choice is not exhaustively explored within the present paper. The diverged top-ten sets in Fig. 9 indicate that the health-safety weight is non-trivially driving

recommendations at the chosen 0.4 setting; a full sensitivity grid over the weight axis is left to future work. Deployment of the framework in production carries obligations that lie outside the technical scope of this paper. Health-condition data is protected under most jurisdictions' privacy law, including the health-information provisions of the European General Data Protection Regulation and equivalent regimes elsewhere. Any production integration would require appropriate consent flows, data-minimization discipline, and secure storage. The recommendation output is itself a sensitive artifact because it reveals inferences about a user's health, and any user-facing surface that displays or persists these recommendations should account for that sensitivity. The ethical case for the framework rests on informed user opt-in and on the framework being described as an information aid rather than a substitute for travel-medicine consultation. CodeCarbon emission estimates depend on the inferred grid mix of the runtime environment, and a single training session does not capture inference-time or full lifecycle emissions. The relative comparison reported here is therefore more robust than the absolute numbers.

6. Conclusions

This paper introduces a framework for travel recommendation that integrates real environmental measurements and medically grounded health-condition profiles into the feature representation, and evaluates eleven machine learning models against the joint objectives of predictive accuracy and training-time carbon emissions. On the present benchmark, XGBoost achieves the lowest RMSE and the highest NDCG@10 while emitting the second-lowest CO₂ in the model family. Across the full Random Forest complexity sweep evaluated here, no configuration matches gradient boosting on either axis. The Pareto frontier of accuracy and sustainability is therefore concentrated on the boosting end of the model space for this task class, a pattern consistent with the wider tabular-data literature and demonstrated in the specific applied context of health-aware travel recommendation. The health-aware feature pipeline shifts top-ten recommendations for the affected subpopulations along clinically interpretable axes, providing an implicit ablation against a preference-only baseline even though global feature-importance averages dilute the signal across the healthy majority. The full replication code is released as an open repository [24]. The present results should be read as a methodological proof of concept; clinical validation on a cohort of real travel-medicine patients, replication at deployment scale with thousands of points of interest, and a formal sensitivity analysis over the preference-versus-health weighting are the principal extensions the framework requires before deployment.

Funding: This article has been funded by the Deanship of Scientific Research in King Faisal University with Grant Number KFU263550.

Acknowledgments: The authors are thankful to King Faisal University for funding and supporting this research.

Data Availability Statement: The full data construction code, the cached environmental features, and the trained model artifacts are released as an open-source repository [24]. The user preference data originates from the publicly available UCI Travel Review Ratings dataset [3] and is redistributed within the replication package under the original license.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3645–3650.
2. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. *Communications of the ACM* 2020, 63, 54–63.
3. Renjith, S.; Sreekumar, A.; Jathavedan, M. Travel Review Ratings Dataset. UCI Machine Learning Repository, 2018. Available online: <https://doi.org/10.24432/C5G62B> (accessed on 16 June 2026).
4. Borràs, J.; Moreno, A.; Valls, A. Intelligent Tourism Recommender Systems: A Survey. *Expert Systems with Applications* 2014, 41, 7370–7389.
5. Yang, D.; Zhang, D.; Zheng, V.W.; Yu, Z. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 2015, 45, 129–142.
6. Cho, E.; Myers, S.A.; Leskovec, J. Friendship and Mobility: User Movement in Location-Based Social Networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 1082–1090.
7. Bondevik, J.N.; Bennin, K.E.; Babur, Ö.; Ersoy, C. A Systematic Review on Food Recommender Systems. *Expert Systems with Applications* 2024, 238, 122166.
8. Luks, A.M.; Hackett, P.H. Medical Conditions and High-Altitude Travel. *New England Journal of Medicine* 2022, 386, 364–373.
9. American College of Obstetricians and Gynecologists. Air Travel During Pregnancy. ACOG Committee Opinion No. 746, 2018, reaffirmed 2024.
10. World Health Organization. WHO Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide; World Health Organization: Geneva, Switzerland, 2021.
11. Henderson, P.; Hu, J.; Romoff, J.; Brunskill, E.; Jurafsky, D.; Pineau, J. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *Journal of Machine Learning Research* 2020, 21, 1–43.
12. Courty, B.; Schmidt, V.; Goyal-Kamal; Coutarel, M.; Feld, B.; Lecourt, J.; Liné, J.; et al. CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing. Zenodo, 2024.
13. Patterson, D.; Gonzalez, J.; Le, Q.; Liang, C.; Munguia, L.M.; Rothchild, D.; So, D.; Texier, M.; Dean, J. Carbon Emissions and Large Neural Network Training. *arXiv preprint arXiv:2104.10350*, 2021.
14. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data? In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, New Orleans, LA, USA, 28 November–9 December 2022; pp. 507–520.
15. World Health Organization. Global Solar UV Index: A Practical Guide; World Health Organization: Geneva, Switzerland, 2002 (reaffirmed in subsequent guidance).
16. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, USA, 4–9 December 2017; pp. 3146–3154.
17. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
18. Breiman, L. Random Forests. *Machine Learning* 2001, 45, 5–32.
19. Wang, S.; Hu, L.; Wang, Y.; He, X.; Sheng, Q.Z.; Orgun, M.A.; Cao, L.; Ricci, F.; Yu, P.S. A Survey on Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions. *ACM Transactions on Recommender Systems* 2024, 1, 1–51.
20. Tran, T.N.T.; Atas, M.; Felfernig, A.; Stettinger, M. An Overview of Recommender Systems in the Healthy Food Domain. *Journal of Intelligent Information Systems* 2018, 50, 501–526.
21. Ricci, F.; Rokach, L.; Shapira, B. (Eds.) *Recommender Systems Handbook*, 3rd ed.; Springer: New York, NY, USA, 2022.
22. Lacoste, A.; Luccioni, A.; Schmidt, V.; Dandres, T. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700*, 2019.
23. Zippenfenig, P. Open-Meteo: Free Open-Source Weather API. Available online: <https://open-meteo.com> (accessed on 16 June 2026).
24. Zia, M.A.; Riaz, S.; Noor, M.; AlJabr, A.; Abual-Rub, M.; Abu-Zanona, M. Health-Aware Green Travel Recommender: Reproducibility Package. Zenodo, <https://doi.org/10.5281/zenodo.20940672> (accessed on 26 June 2026).
25. Verma, A.; Patel, S.; Kumar, R. A Comprehensive Review of Green AI Techniques for Sustainable Machine Learning. *Sustainable Computing: Informatics and Systems* 2024, 41, 100945.
26. García-Martín, E.; Rodrigues, C.F.; Riley, G.; Grahn, H. Estimation of Energy Consumption in Machine Learning. *Journal of Parallel and Distributed Computing* 2019, 134, 75–88.
27. Berthelot, A.; Caron, E.; Jouvelot, P. Estimating the Environmental Impact of Generative AI Services: A Systematic Methodology. *Future Generation Computer Systems* 2024, 158, 213–225.
28. Ie, E.; Hsu, C.-W.; Mladenov, M.; Jain, V.; Narvekar, S.; Wang, J.; Wu, R.; Boutilier, C. RecSim: A Configurable Simulation Platform for Recommender Systems. *arXiv preprint arXiv:1909.04847*, 2019.