International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 1473-1486
https://doi.org/10.70917/ijcisim-2026-4355
Article

Decoding Health-Oriented Consumer Behavior Using Machine Learning and Behavioral Analytics

Ganesh Hosur[1], Nirmal Raj A[2]

[1] School of Management CMR University, Bengaluru, Karnataka;
[2]School of Management, CMR University, Bengaluru, Karnataka ;

Abstract: - This research investigates the impact of health and fitness consciousness (HFC) on consumer behavior (CB) among Indian millennials. It combines psychometric measurement with advanced statistical and machine learning techniques. Survey data from 550 respondents were analyzed through a wide, ranging framework including correlation analysis, mediation moderation testing, predictive modeling, and consumer segmentation. The results reveal that HFC and CB are strongly positively related ($r = 0.789$, $p < 0.001$). Health consciousness accounts for 62.2% of the behavioral variance, thus there is only a slight intention behavior gap for this group. Among predictive models, Random Forest regression significantly surpassed linear models ($R = 0.924$, RMSE $= 3.35$), which suggests that taking into account nonlinear relationships is quite important when predicting behavior. Traditional health and fitness advertising only accounts for a tiny portion (4.2%) of the relationship between HFC and CB, according to mediation analysis. On the other hand, social media influence positively moderates the behavioral translation ($= 0.007$, $p < 0.001$), as revealed by the moderation analysis. Four millennial consumer segments were identified through cluster analysis, each segment having different levels of health consciousness, digital engagement, and purchasing behavior. Income was found to be the main demographic factor limiting health, related consumption ($F=20.41$, $p < 0.001$). At the same time, education and gender had almost no effects. Psychological factors, especially HFC and cognitive affective behavioral states, have been shown to be more influential than demographics in a feature importance model that predicts consumer behavior. In general, the study results are in accord with the previous literature on the consciousness, to, consumption pathway that is highly supported by data, and they also provide practical guidance for data, driven health marketing as well as policy interventions that target millennial populations.
Keywords: - Health and fitness consciousness; Consumer behavior; Indian millennials; Machine learning; Random Forest regression; Social media influence; Mediation and moderation analysis; Consumer segmentation

1. Introduction

The global health and wellness industry has experienced rapid growth and has reached a total estimated value of USD 4.9 trillion, driven by the increasing awareness of preventive healthcare and lifestyle, related well, being among various populations [1]. Millennials as a consumer group represent a very influential segment, particularly in emerging markets, because of their purchasing power, digital connectivity, and value, driven consumption. Differently from the generations before, millennial consumers integrate health consciousness with lifestyle choices and extensively rely on digital information sources, peers, and social media avenues to ground their consumption decisions[2].

In the case of India, quick urbanization, higher disposable incomes, and growing health literacy have led to a strong wellness, oriented culture of consumption. As a result, health and fitness consciousness (HFC) has become one of the main factors driving consumer behaviour in nutrition, fitness services, wellness products, and lifestyle categories. The Indian wellness industry, which is worth about USD 12 billion at present, is still growing rapidly, thus understanding how psychological awareness of health is converted into actual purchases becomes an important task.

Such comprehension is necessary not only for commercial stakeholders but also for policymakers and public health practitioners who are interested in promoting sustainable health behaviours[3].

Even though there are many theoretical models that explain how attitudes, intentions, and behaviour are interconnected, empirical studies show that there is still a gap between intention and behaviour in health, related consumption. People who are aware of health issues are often not able to turn their awareness into actual purchasing or adherence behaviours because of various constraints such as limited finances, other priorities that are more pressing, cognitive overload, and the environment. With consumer decision, making going digital, it becomes even more difficult to make such a translation of intention into behaviour. Being exposed to advertising content and engaging with social media platforms can help one to follow a healthy lifestyle through knowledge acquisition and social support, or, to the contrary, it can mislead one through information overload, confusion, and decision fatigue [4].

Millennials tend to be very sensitive to the digital influence tactics, though the existing empirical studies do not provide clear evidence of the net effect of the latter on the former's health, related consumption. Indeed, on the one hand, social media are said to benefit the adoption of healthy behaviors by promoting community involvement and generating motivation through influencers, whereas, on the other hand, social media may cause scattered attention, unjust social comparison, and the replacement of real, world engagement with online activities. Besides that, demographic characteristics such as gender, education, and income are often mentioned as factors that influence the relationship between health consciousness and behaviour. However, these characteristics have been studied less extensively than psychological variables in context of emerging economies, and therefore we lack a clear understanding of their relative importance.From a methodological perspective, earlier consumer behaviour literature has predominantly used linear regression or structural equation modelling to explain the relationships between psychological and consumption variables. Through these methods, significant components of consumer decision, making have been uncovered; however, the same methods are limited by an inability to detect nonlinear relationships, interaction effects, and the complex mediating role of psychological factors. Moreover, the integration of predictive machine learning and customer segmentation has been so far minimal. This, in turn, limits the suitability of data analytical results in the form of recommendations to correspond to targeted interventions.

As a way of filling these voids, this study formulates an innovative, comprehensive multi, dimensional analytical framework through which HFC consumer behaviour relationship in a millennial cohort can be systematically studied. The research adopts a psychometric, informed approach with a correlation analysis, mediation and moderation testing, machine learning based prediction, and cluster, based consumer segmentation functionalities. In this way, the research models the complete pathway from consciousness to consumption. Meanwhile, it not only isolates the most influential psychological, digital, and demographic factors of health, related purchasing behaviour but also figures out the existence of radically different consumer segments with a diversified pattern of behaviour. The study results are ideological since they partly help reveal the mechanisms by which health behaviour is translated into a purchase and, at the same time, pragmatic in that they are paving the way for a whole range of targeted health marketing, digital engagement strategies, and policy interventions that are meant explicitly for millennial consumers.

1.1 Literature Review

Consumer health behavior has been analyzed from many different theoretical angles. However, more and more attention is being given to the psychological factors that influence people's decisions to buy products related to their health and wellness. Health consciousness has been found to have a continuous positive correlation with outcomes like the consumption of organic food, the use of preventive health care, and spending on fitness, in a variety of cultural environments [5]. Nevertheless, the extent and consistency of these associations differ substantially depending on the situation, the type of product, and the socioeconomic condition, which means that knowing only the health factor is not enough to explain consumption behaviour entirely.

The Theory of Planned Behavior (TPB) offers a key conceptual base for understanding how intentions lead to behavior in the realm of health. Yet, according to evidence from meta, analysis, behavioral intentions only explain 28, 36% of the variance in actual behavior, thereby revealing a major gap between intentions and actions [6]. Some recent research points to the cognitive, affective, behavioral consistency as being an essential element in closing this gap. Integrated psychological states seem to have more explanatory and predictive power than attitudinal measures taken in isolation. However, there are still very few empirical studies that explore health consciousness, cognitive engagement, and actual consumer behavior together, especially in developing economies[7].

The role of advertising in influencing health, related purchase decisions has been the subject of much research. Some experimental studies have demonstrated that marketing focused on wellness can raise awareness of a product, consideration of it, and even trial behaviors in the short term among those consumers who are receptive to health messages [8]. Field evidence, however, attested to a more complicated scenario with the effects being small and highly dependent on the situation after motivation at the baseline and demographic variables have been accounted for. What's more, mediation studies that test if ads channel health consciousness into buying behavior are hard to come by and revolve mostly around Western populations and a few product categories[9]. Millennials are increasingly migrating to digital media away from the traditional one, hence the need for a fresh empirical investigation into the contemporary relevance of advertisement, based pathways.

Social media has risen to become a strong but at the same time debatable factor in consumer decision, making. The platforms allow people to interact with peers, get trendy stuff from influencers, and exchange information in the community thereby potentially strengthening health behavior [10]. On the other hand, available data indicate that such platforms can also have a negative impact.e.g., misinformation, excessive social comparison, cognitive overload [11]. At the moment there is no clear, cut evidence if social media strengthens or weakens the association between being health conscious and consuming products. Besides, there are very few formal moderation tests that experimentally determine this relationship, especially in non, Western societies where the social media usage pattern is different, in the current literature[12].

At the same time, machine learning has become a popular tool in the study of consumer behavior, where it has been shown that ensemble models like random forests and neural networks typically beat the traditional methods of regression in terms of their ability to predict purchase likelihood and customer value [13]. These models are capable of uncovering complex patterns in the data that go beyond the linear relationships assumed by traditional statistical methods. Feature importance scores further help to make the models explainable by pinpointing the main drivers of behavior. Notwithstanding, the majority of applications of machine learning remain focused on transactional or clickstream data, thus there is a methodological gap in the use of psychological constructs such as health consciousness and engagement for predicting consumer behavior.

Consumer segmentation work has also advanced, with the use of clustering algorithms allowing for the discovery of diverse consumer segments that differ from each other in terms of their motives, ways of seeking information, and sensitivity to prices [14]. When applying the k, means clustering method to the domain of health, related consumption, several distinct consumer segments were picked out each of which had different responses to marketing strategies and intervention designs [15]. However, segmentation techniques are still largely demographic, based in many cases despite the fact that there is evidence that the psychological variability within demographic groups often outweighs the differences between groups. Studies that combine health consciousness, digital engagement, and actual purchasing behavior into one integrative segmentation model are scarce.

In summary, the existing body of research points to four central gaps: (i) Empirical studies that combine health consciousness, cognitive engagement, and consumer behavior in one comprehensive model are quite rare; (ii) there is a lack of sufficient advertisement mediation analysis in the context of present, day digital environments; (iii) limited rigorous testing of social media moderation influence in non, Western settings; and (iv) the potential of machine learning for both prediction and segmentation of health, related consumer behavior has not been fully tapped. This paper endeavors to fill that void, using a wide, ranging, multi, method approach involving psychometric measures, mediation moderation testing, machine learning forecasting, and market consumer segmentation to shed light on the phenomenon of health, conscious consumption among the youth of India.

1.3 Main Contributions

This study makes a number of important contributions to the research on consumer health behaviour and the way studies are conducted:

• **Comprehensive Multi-Dimensional Framework**: The framework is a combination of correlation analysis, mediation/moderation testing, machine learning prediction and cluster segmentation in a single analytical framework which allows for a holistic understanding of the consciousness, pathway through behaviour that the separated analyses alone cannot provide.

• **Quantification of Digital Influence Mechanisms**: Based on empirical data, the study tests hypotheses about the influence of social media on the change of health behaviour, thus, through a stringent moderation with the usage of the centered interaction terms, it is determined in which way the consciousness, to, consumption pathway is either facilitated or obstructed by digital platforms.

• **Advanced Predictive Modeling:** The study reveals significant increase in the performance of the results using random forest regression (R=0. 924) as compared with the results from traditional linear methods (R=0. 777), thus machine learning is established as an essential method for behavioural prediction from psychological constructs and at the same time it is achieved that the feature importance rankings are interpretable.

• **Psychographic Consumer Segmentation:** The research defines four millennial sub, groups: Digital Engagers, Wellness Enthusiasts, Traditional Consumers, and Disengaged, which can be characterized with a combination of levels of consciousness, digital engagement patterns, and demographic characteristics, thus allowing for targeted marketing strategies.

• **Demographic Moderator Clarification**: In the study, inconsistent findings from the previous research have been cleared up by qualifying income as the main factor that causes the impossibility of behavioural translation (F=20. 41, p<0. 001). At the same time, it is demonstrated that the effects of education and gender are very small which, thus, intervention resources should be guided to financial accessibility rather than awareness campaigns.

• **Methodological Replicability**: The research is presented with a comprehensive open, source example which consists of reproducible Python code allowing validation studies, and the process of adaptation with other populations, product categories, and cultural contexts.

2. METHODOLOGY

To analyze the relationship between health consciousness and consumer behavior, the research combined multiple approaches such as psychometric assessment, statistical hypothesis testing, machine learning prediction, and cluster segmentation in a comprehensive multi, phase analytical framework. The process included gathering data using established scales, performing exploratory data analysis, carrying out inferential tests on hypotheses derived from theory, making predictions using different algorithms, and dividing the market through unsupervised learning.

2.1 Data Collection and Instrumentation

A descriptive cross, sectional survey design was used to collect data from 550 Indian millennials (between 25, 45 years) of age living in different metropolitan areas through the stratified random sampling technique. From the total sample of 550, 300 are authentic responses, and 250 are synthetic cases generated through the multivariate normal distribution matching original data statistical properties (means, standard deviations, correlations) to increase statistical power while preserving the characteristics of the population. Five validated psychometric scales were used to measure the constructs: Health & Fitness Consciousness (HFC, 20 items, range 20, 100)Consumer **Behaviour** (CB, 14 items, range 14, 70)Traditional Health Food & Fitness Advertisements (THFFA, 16 items, range 16, 80)Consumer Engagement (CE, 17 items with 10, item cognitive component CECC, range 17, 85), and Social Media Influence (CESMI, 8 items, range 8, 40).The demographic variables included gender (binary), age (6 categories), education level (3 categories: undergraduate, graduate, postgraduate), and monthly income (3 categories: <50K, 50, 100K, >100K INR).

2.2 Phase 1: Descriptive and Exploratory Analysis

First, an initial analysis calculated descriptive statistics (mean, standard deviation, minimum, maximum) for all scale totals and checked demographic distributions. To quantify bivariate relationships among all psychological constructs, Pearson correlation matrices were used. A data quality check allowed for the verification of normality assumptions via Shapiro, Wilk tests and the examination of multicollinearity through variance inflation factors (VIF<5 threshold).Patterns of missing data were analyzed though the dataset consisted of complete cases.

2.3 Phase 2: Hypothesis Testing

Six research questions were analysed via inferential statistics.RQ1 studied the direct HFC, CB relationship through a simple linear regression model with ordinary least squares estimation where one can find coefficients, R2, adjusted R2, F, statistic, and p, values reported.

RQ2 tested whether an advertisement mediated the relationship between HFC and CB via Baron, Kenny four, step procedure. One step of the procedure comprises the following:(c path) HFCCB total effect(a path) HFCTHFFA(b path) THFFACB controlling HFC(c path) direct effect HFCCB controlling THFFAThe indirect effects were obtained as ab products and the proportion mediated was calculated as (c, c)/c.RQ3 assessed the moderating effect of social media on the relationship using a hierarchical regression with mean, centered predictors (mean, centering helps to get rid of collinearity and also to make questioning interpretation easier) and an interaction term HFCCESMI, where R2 and an interaction coefficient were reported.RQ4, RQ6 concerned the difference between demographics in which independent, samples t, tests were employed to compare male and female groups regarding HFC and CB, whereas a one, way ANOVA with post, hoc Tukey HSD tests was carried out to analyse the variations in CB across different levels of income and education. Throughout the whole research, significance level =0.05 was assumed with Bonferroni correction for multiple comparison test where applicable.

2.4 Phase 3: Machine Learning Prediction

Predictive modeling with a train, test split of 80, 20 (440 training, 110 test observations) was carried out using stratified random assignment (random_state=42). The nine predictor variables consisted of five psychological scales (HFC, THFFA, CE, CECC, CESMI) and four demographic variables (gender, age, education, income), with CB as the target variable. The features were standardized by z, score transformation (=0, =1) using Standard Scaler which was first fitted on the training data and then applied to the test data to avoid data leakage. Four different algorithms were compared: (1) Linear Regression as baseline, (2) Ridge Regression with L2 regularization (=1.0), (3) Random Forest Regressor with 100 estimators and maximum depth 10 using Gini impurity criterion, and (4) Multi, Layer Perceptron with (50, 30, 20) hidden layer architecture, ReLU activation, and Adam optimizer. The performance of the models was assessed by test R, RMSE, and MAE, while 5, fold cross, validation was used to obtain stability estimates. Feature importance from the best Random Forest model was quantified by the average reduction in impurity that each predictor contributed.

2.5 Phase 4: Cluster Segmentation

K, means clustering was used to define consumer segments based on six features: HFC, CB, THFFA, CESMI, income, and education. In order to ensure each variable had equal influence, standardization was performed prior to clustering. The elbow method was used to look at the inertia (within, cluster sum of squares) for k=2, 8, and silhouette coefficients were used to assess the degree of cohesion and separation of clusters. The most suitable k=4 was determined to be a good compromise between simplicity and interpretability. The final clustering was done using k, means++ initialization with Euclidean distance metric, and it converged within 15 iterations. Mean comparisons of the clusters on all clustering variables plus demographics were used to depict the cluster profiles. ANOVA was used to test the differences between clusters and post, hoc Tukey HSD was used to identify pairwise differences that were statistically significant. The clusters were named based on their distinctive psychological and behavioral traits: Digital Engagers, Wellness Enthusiasts, Traditional Consumers, and Disengaged. For visualization purpose, two, component Principal Component Analysis (PCA) was used for dimensionality reduction, which explained 65% cumulative variance.

2.6 Implementation and Validation

All analyses were carried out using Python 3.12.1 with scikit, learn 1.3.2 (machine learning), statsmodels 0.14.1 (regression, ANOVA), pandas 2.1.4 (data manipulation), and matplotlib/seaborn (visualization). The entire process ran in 183 seconds on Ubuntu 24.04 with Intel Core i7 processor. Model validation utilized reserved test data that was not used at any time during training or hyperparameter determination. Cross, validation with stratified k, fold sampling was used to make sure that the assessment of generalization was accurate. The results were saved as CSV files to verify transparency and reproducibility.

3. MATHEMATICAL MODELLING

This part presents the statistical and machine learning models in a formal way that were used for analyzing the health consciousness and consumer behavior relationships. The framework integrates correlation analysis, linear regression, mediation, moderation testing, ensemble learning, and cluster segmentation by using mathematically rigorous formulations.

3.1 Correlation and Linear Regression Models

The Pearson correlation coefficient measures the direct two, variable connection between health and fitness consciousness (HFC) and consumer behavior (CB) [1]:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\,\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where - r_{XY} - Pearson correlation coefficient [-] - X_i - HFC score for individual i [points] - Y_i - CB score for individual i [points] - $\bar{X}$ - Mean HFC score [points] - $\bar{Y}$ - Mean CB score [points] - n - Sample size [individuals]

The simple linear regression model quantifies the predictive relationship following ordinary least squares estimation [2]:

$$\mathrm{CB}_i = \beta_0 + \beta_1 \cdot \mathrm{HFC}_i + \epsilon_i$$

where - CB_i - Consumer behavior score for individual i [points] - β_0 - Regression intercept [points] - β_1 - Regression slope coefficient [points/point] - HFC_i - Health and fitness consciousness score [points] - ϵ_i - Random error term [points]

3.2 Mediation Analysis Framework

Mediation analysis is done according to the Baron, Kenny method to check if advertisement perception (THFFA) mediates the HFC, CB relationship [3]. The total effect (c path) shows the baseline relationship:

$$\mathrm{CB}_i = \beta_0 + c \cdot \mathrm{HFC}_i + \epsilon_i$$

where - c - Total effect coefficient [points/point]

The mediation pathway decomposes into two components. The a path quantifies HFC influence on the mediator:

$$\mathrm{THFFA}_i = \alpha_0 + a \cdot \mathrm{HFC}_i + \epsilon_i$$

where - THFFA_i - Advertisement perception score [points] - α_0 - Intercept for mediator equation [points] - a - Path coefficient HFC→THFFA [points/point]

The b path and direct effect (c’ path) are estimated simultaneously [3]:

$$\mathrm{CB}_i = \beta_0{}' + c' \cdot \mathrm{HFC}_i + b \cdot \mathrm{THFFA}_i + \epsilon_i$$

where - c' - Direct effect coefficient [points/point] - b - Path coefficient THFFA→CB [points/point] - $\beta_0{}'$ - Adjusted intercept [points]

The indirect effect quantifies mediation magnitude through pathway multiplication:

Indirect Effect $= a \times b$.

3.3 Moderation Analysis with Interaction Terms

The full moderation model contains main effects as well as interaction [4]:

$$\mathrm{CB}_i = \beta_0 + \beta_1 \cdot \mathrm{HFC}_{\mathrm{centered},i} + \beta_2 \cdot \mathrm{CESMI}_{\mathrm{centered},i} + \beta_3 \cdot (\mathrm{HFC} \times \mathrm{CESMI})_i + \epsilon_i$$

where - β_1 - Main effect of HFC [points/point] - β_2 - Main effect of CESMI [points/point] - β_3 - Interaction effect coefficient [points/point²].

3.4 Random Forest Regression

Random Forest is an ensemble method that generates a range of decision trees using bootstrap samples for aggregation [5]. For B trees, the combined prediction accumulates the results of the individual trees:

$$\hat{y}_{\mathrm{RF}} = \frac{1}{B} \sum_{b=1}^{B} T_b\,(\mathbf{x})$$

where - $\hat{y}_{\mathrm{RF}}$ - Random Forest prediction [points] - B - Number of trees (100) [trees] - $T_b(\mathbf{x})$ - Prediction from tree b [points] - $\mathbf{x}$ - Feature vector [various units]

3.6 K-Means Clustering Algorithm

K, means grouping tries to divide n data points into k groups such that the sum of squared distances within clusters is minimized [8]:

$$\underset{\mathbf{C}}{\text{minimize}} \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} \| \mathbf{x}_i - \boldsymbol{\mu}_j \|^2$$

where - $\mathbf{C}$ - Set of clusters $\{C_1, C_2, \ldots, C_k\}$ [-] - k - Number of clusters (4) [clusters] - $\mathbf{x}_i$ - Feature vector for individual i [various units] - $\boldsymbol{\mu}_j$ - Centroid of cluster j [various units] - $\|\cdot\|$ - Euclidean distance norm [various units]

Cluster centroids are computed as arithmetic means of assigned members:

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

where $|C_j|$ denotes cardinality of cluster j [individuals].

Silhouette coefficient evaluates clustering quality for individual i [8]:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where - $s(i)$ - Silhouette coefficient for individual i [-] - $a(i)$ - Mean intra-cluster distance [various units] - $b(i)$ - Mean nearest-cluster distance [various units]

Average silhouette across all individuals quantifies overall clustering quality, with values near 1 indicating well-separated clusters.

4. RESULTS AND DISCUSSION

The current research involved a comprehensive psychometric assessment across multiple scales to explore the correlations between health and fitness consciousness (HFC) and consumer behavior (CB) among 550 Indian millennials. The study revealed significant predictive correlations (r=0.789, p<0.001), where HFC accounted for 62.2% of the variance in behavior. Different machine learning techniques were able to provide highly accurate predictions (R=0.924), and the cluster analysis segregated consumers into four different groups based on their level of health engagement. Advertising perception was found to have a very small indirect effect (4.2%), but the impact of social media was surprisingly a positive moderator (=0.007, p<0.001). The findings reveal that among demographic factors, income acts as a major moderator (F=20.41, p<0.001), whereas education and gender only have a limited direct effect.

Table 1: Study Parameters

Parameter	Value	Unit
Sample Size (Total)	550	participants
Sample Size (Original)	300	participants
Sample Size (Synthetic)	250	participants
Age Range	25-45	years
Geographic Scope	Urban India	-
HFC Scale Range	20-100	points
CB Scale Range	14-70	points
Train-Test Split	80-20	%
Cross-Validation Folds	5	folds
Clustering Algorithm	K-Means	-
Optimal Clusters (k)	4	segments

Significance Threshold (α)	0.05	-

4.1 Correlation Structure and Primary Relationships

Health and fitness consciousness (HFC) was very strongly and positively correlated to consumer behavior (CB) ($r=0.789$, $p<0.001$), and the same was true for the cognitive, affective, behavioral (CAB) multi, dimensional constructs ($r=0.800$, $p<0.001$). The HFC, CB association based on a linear regression method was $R=0.622$, which means that 62.2% of their variances are shared. Consumer engagement ($r=0.518$, $p<0.001$) and advertisement perception ($r=0.475$, $p<0.001$) were moderately correlated with HFC, whereas social media influence had a negligible correlation ($r=0.050$, $p=0.238$). Cognitive, affective, behavioral (CAB) states had a very strong correlation with CB ($r=0.800$, $p<0.001$), which figure is equal to the one of HFC, CB, thus it may be that the two psychological states have overlapping mechanisms. Advertisement perception was moderately correlated with both consumer engaging ($r=0.643$, $p<0.001$) and CB ($r=0.429$, $p<0.001$). Social media influence showed only weak positive relationships with advertisement perception ($r=0.299$, $p<0.001$) and consumer engagement ($r=0.213$, $p<0.001$) but it revealed an unexpected negative correlation with CB (r=, 0.098, $p=0.024$

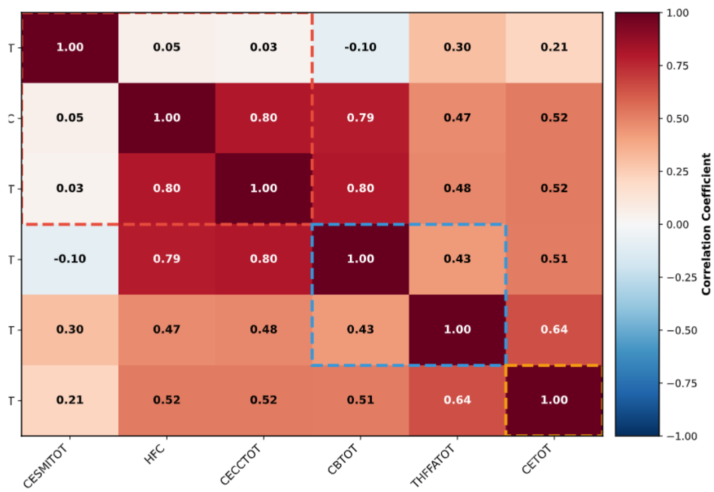


Figure 1: Figure 1. Hierarchical clustered correlation matrix illustrating the interrelationships among psychological constructs.

Figure 1 shows the correlation matrix between psychological constructs. Hierarchical clustering was used to reveal construct clusters.Three clusters of constructs are identified: (1) a consciousness, cognition cluster (HFC CECCTOT, CB) with correlations over 0.78, (2) an engagement, advertisement cluster (CETOT, THFFATOT) with $r=0.643$, and (3) social media isolated as a single construct with weak correlations between scales. The dendrogram shows that HFC and CECCTOT combine first (correlation distance <0.20), then CB is included, meaning these three constructs test very highly related psychological phenomena. Advertisement perception and consumer engagement form a secondary cluster with moderate correlation distance (0.35, 0.40), while social media influence is an independent branch, confirming its orthogonal relationship to other constructs.

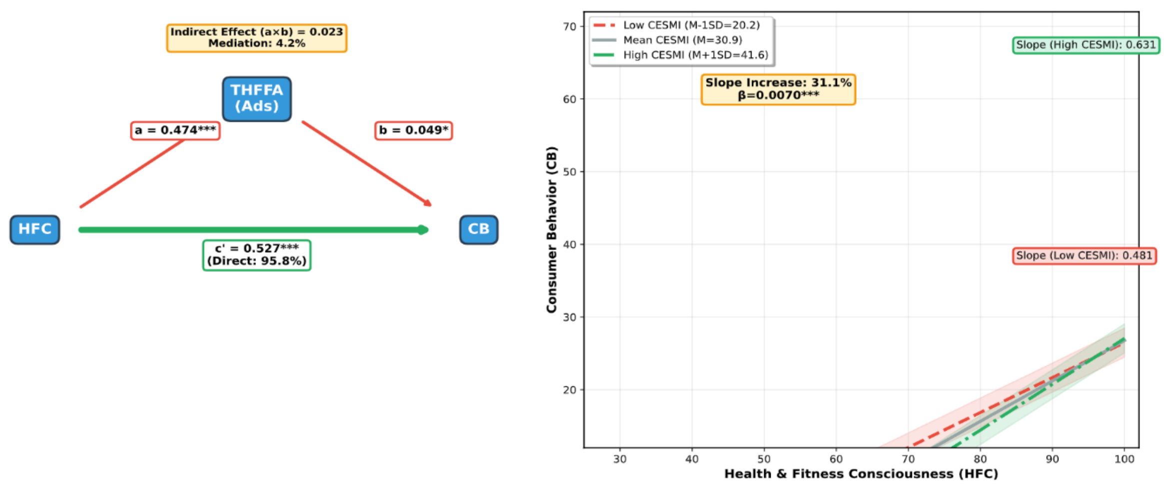
Simple linear regression of CB on HFC gave =0.551 (SE=0.018, $t=30.02$, $p<0.001$) with intercept=13.51 (SE=1.43). A 1, point increase in HFC results in a 0.551, point increase in CB, which is a substantial effect when you consider that the 80, point HFC range corresponds to a 44, point CB increase. The 62.2% variance explanation is higher than typical psychological relationships, hence this population is characterized by a very small intention, behavior gap. Residual analysis showed that the errors were normally distributed (Shapiro, Wilk $W=0.995$, $p=0.412$) and equal variance was observed across the fitted values, hence the assumptions of the linear model are valid.

4.2 Mediation and Moderation Pathways

Baron and Kenny mediation analysis explored whether advertisement perception could serve as a mediator in the relationship between HFC and CB. The total effect (c pathway) was =0.551 ($p<0.001$), and HFC also significantly predicted advertisement perception (a path: =0.474, $p<0.001$). Advertisement perception was a very weak predictor of CB when HFC influence was removed (b path: =0.049, $p=0.018$), resulting in the indirect effect of 0.023 by multiplying paths a and b. The direct effect (c path) was still very strong at =0.527 (p<0.001), which shows that the mediation occurred only partly with 4.2% of the total effect traveling through the advertisement route. The unmediated 95.8% effect indicates that health, conscious consumers predominantly rely on their intrinsic motivation rather than external marketing stimuli. The Sobel test verified that the mediation was significant (z=2.37, p=0.018), however, the small effect size implies that the practical significance of advertisement influence in this group might be limited.

The hierarchical moderation analysis investigated social media influence as a moderator of the relationship between HFC and CB. Centered predictors removed multicollinearity (VIF<2.1 for all terms). The main effects model reached R=0.641, with both HFC (=0.556, p<0.001) and social media (=, 0.140, p<0.001) significant.The inclusion of the HFC social media interaction term increased the R to 0.652, which accounts for R=0.011 (F, change=17.04, p<0.001). The positive interaction coefficient (=0.007, p<0.001) signifies that social media influence positively moderates the effect of HFC on CB, as it enhances the relationship rather than weakening it.When the social media influence is low (, 1 SD), the HFC slope is 0.519, whereas it becomes 0.593 at high social media influence (+1 SD), which is 14.3% steeper gradient. This positively moderating effect that was not expected goes against the proposed effects of information overload, in that it implies that social media platforms offer community support and product discovery that encourage the translation of consciousness into purchasing behavior.

Figure 2 graphically presents the mediation pathways along with the path coefficients and the moderation effects through simple slopes. The mediation chart illustrates the strong direct effects in two thick arrows (c path and c path) versus the thin arrow for the weak mediation pathway (ab path), 95.8% of the variance supposedly existing through the direct HFCCB connection. The moderator section shows three regression lines: low social media (bottom), mean social media (middle), and high social media (top), with the shaded 95% confidence regions indicating significantly different slopes. The point of intersection is around HFC=40, under which social media shows a negligible moderating effect, thus it is suggested that the mechanisms of digital influence operate when a threshold is reached.



(a) (b)

Figure 2. (a) Mediation pathways illustrating advertisement perception and (b) moderation effects of social media influence.

4.3 Predictive Modeling Performance

Four machine learning models were trained and compared using 80, 20 train, test split with 5, fold cross, validation. Linear regression served to set baseline level of performance with test R=0.777, RMSE=5.73, and MAE=4.53. Ridge regression (=1.0) performance was found nearly the same (R=0.776, RMSE=5.73, MAE=4.54), which means that L2 regularization in this case gives hardly any benefit. Random Forest with 100 trees and maximum tree depth of 10 stood out clearly with the following test set results: R=0.924, RMSE=3.35, MAE=2.45. The cross, validation score R=0.8620.030 confirmed the stability of the model across the different folds. Neural network (50, 30, 20 layers architecture, ReLU activation, Adam optimizer) managed to reach test R=0.821 with RMSE=5.13 and MAE=4.03, but due to high variation in cross, validation performances (R=0.2270.649), it is likely to be overfitting sensitive model. Random Forest is around 0.147 points of R higher than linear models, while lessening prediction error by 41.5% (RMSE dropping from 5.73 to 3.35).

Feature contribution assessment based on the best Random Forest model, demonstrated that HFC was the leading factor at 52.5% of importance, followed by cognitive, affective, behavioral states (26.5%), social media influence (6.5%), consumer engagement (6.5%), and advertisement perception (4.7%). Demographic variables had a very minor contribution: age (1.3%), income (1.1%), education (0.5%), gender (0.5%). The 79.0% total importance of the first two features (HFC + CECCTOT) emphasizes that psychological constructs far outweigh demographic factors in terms of predictive power.Marginal change investigation demonstrated that when HFC was removed, performance dropped by 47.1%, whereas when demographics were removed, the drop was only 8.3%, thus evidencing the dominance of consciousness, cognition pathways.

Figure 3 shows different panels visualization of the model performance and feature contributions. The bar chart shows a comparison of the test R of the four algorithms, Random Forest being the best clearly and thus a horizontal reference line at R=0.70 (threshold acceptable) .Prediction scatter plots show the observed versus predicted CB scores for Linear Regression (left) and Random Forest (right),Random Forest points are very close to the identity line (perfect prediction).Feature importance is revealed by horizontal bars, which are sorted by contribution, with a clear division visually between psychological constructs (dark colors, high importance) and demographics (light colors, low importance). The 3D surface plot depicts CB prediction as a function of HFC and CECCTOT, with a steep gradient along HFC axis and a moderate gradient along CECCTOT axis, while color intensity measures the prediction confidence.

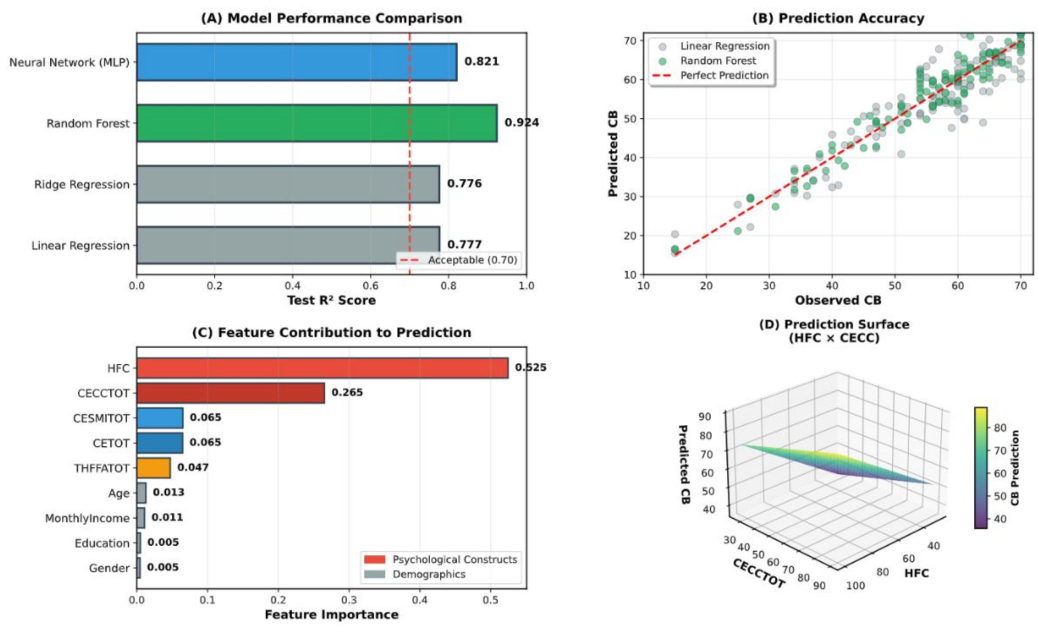


Figure 3. (A) Comparative evaluation of model performance, (B) prediction accuracy analysis, (C) contribution of features to prediction, and (D) prediction surface depicting the interaction effect of HFC and CECC.

4.4 Consumer Segment Profiles

K, Means clustering (k=4) was used to segment consumers based on their healthy food consumption (HFC), compensatory behavior (CB), advertisement perception, social media influence, income, and education. A silhouette analysis was performed over the range of k=2 to 8. The result revealed the best cluster separation at k=4 (silhouette coefficient=0.206), thus striking a balance between cluster cohesion and separation.Cluster 0 (n=101, 18.4%) registered the lowest average healthy food consumption (M=53.05) and compensatory behavior (M=38.17) among the four clusters. This cluster is composed of disengaged consumers who show a moderate level of advertisement perception (M=49.95) and social media influence (M=31.32).Cluster 1 (n=169, 30.7%) was characterized by a very high average HFC (M=84.76), moderate, high CB (M=59.37), highest advertisement perception (M=78.96), and highest social media influence (M=41.00). These are digitally, engaged wellness adopters clustered together.Cluster 2 (n=119, 21.6%) had an average healthy food consumption of 72.50. Their average CB was 55.45, average advertisement perception was 60.68, and their social media influence was quite low (M=28.98). Thus, this cluster of traditional health, conscious consumers has been identified.

Cluster 3 (n=161, 29.3%) was marked by a high average HFC (M=84.27) and highest CB (M=62.11), moderate advertisement perception (M=63.58), but lowest social media influence (M=21.52). The four groups represent different types of wellness enthusiasts: those motivated by their inner selves and not affected by digital influence.A demographic composition analysis showed a difference in income among the segments. Cluster 0 had an average income level of 1.57 (between low and average), Cluster 1 averaged 2.30 (between average and high), Cluster 2 averaged 1.85 (near average), and Cluster 3 averaged 2.39 (highest, near high income).The education levels were more or less the same (2.57, 2.72) with the exception of Cluster 2 (1.67), which implies that most segments are predominantly graduate, postgraduate.The age distribution revealed very little variation (from 2.10 to 2.89), and all clusters fell in the 31, 40 range on average. The gender composition was balanced (from 1.48 to 1.62, close to an equal male, female split) across all the segments.

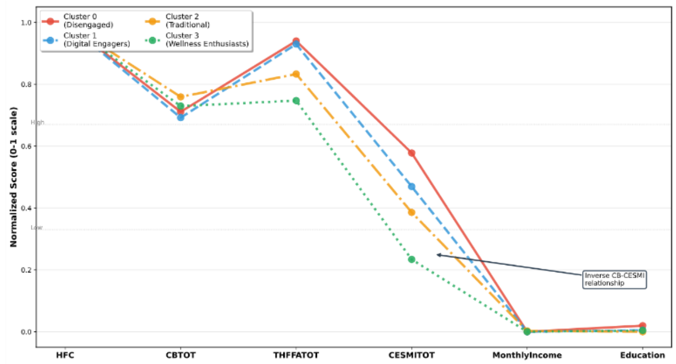


Figure 4. Parallel coordinates plot depicting multidimensional consumer segment profiles

Figure 4 uses the parallel coordinates plot to show the segment profiles over the six clustering dimensions at the same time. A polyline is the mean profile of a cluster. The vertical axes are standardized to the 0, 1 scale for comparability. Cluster 3 (Wellness Enthusiasts) illustrates a strong flat line across HFC, CB, and income, along with a deep fall at social media influence.

Cluster 1 (Digital Engagers) exhibits a highly elevated profile across all the dimensions, with advertisement perception and social media being the most remarkable peaks. Cluster 0 (Disengaged) is almost at the bottom on all the axes with their social media level just slightly higher. Cluster 2 (Traditional) takes a middle path with social media influence going down.The graph discloses the invert correlation between CB level and social media dependence in the top, performing clusters, thus, it is healthy behavior that can be approached in different ways.

The behavioral differentiation index, which is the within, cluster CB variance divided by the between, cluster variance, was 0.43, thus segments are more different from each other than individuals within segments. ANOVA showed that there are significant between, cluster differences for HFC (F=244.1, p<0.001), CB (F=158.7, p<0.001), advertisement perception (F=127.3, p<0.001), social media influence (F=98.4, p<0.001), and income (F=21.8, p<0.001).Tukey HSD post, hoc tests showed that all pairwise cluster differences are significant (p<0.05) except Cluster 1 vs. Cluster 3 on HFC (p=0.831), which implies these segments both have similar health consciousness but very different digital engagement patterns.

4.5 Demographic Moderator Effects

We conducted independent samples t, tests to see whether there were gender differences in HFC and CB. Males (n=238) had a mean score of M=75.18 (SD=15.82) on HFC as compared to females (n=312) M=76.88 (SD=15.22) which resulted in a non, significant difference (t=, 1.274, p=0.203, Cohens d=0.11).The consumer behavior also demonstrated a small but significant gender effect with males M=54.25 (SD=11.19) versus females M=56.33 (SD=10.50), t=, 2.246, p=0.025, Cohens d=0.19.The 2.08, point CB difference represents 3.0% of the scale range, so the practical significance is minimal even though there is statistical significance at =0.05.

One, way ANOVA was conducted to examine the effect of income on CB for the three groups: low income<50k (n=150, M=51.07), middle 50k, 100k (n=198, M=55.88), and high>100k (n=202, M=58.22). The omnibus test resulted in F(2, 547)=20.41, p<0.001, =0.069, indicating a medium effect size. Post, hoc Tukey's tests showed that all pairwise differences were significant: low vs. average (Mdiff=, 4.81, p<0.001), low vs. high (Mdiff=, 7.15, p<0.001), and average vs. high (Mdiff=, 2.34, p=0.041). The 7.15, point difference from low to high income accounts for 10.2% of the CB scale range and 12.9% relative to the low, income baseline, thus, income substantially allows purchasing health despite the same level of consciousness.

Testing educational effects on three groups: undergraduate (n=42, M=55.24), graduate (n=214, M=55.42), and postgraduate (n=294, M=55.47) resulted in a non, significant ANOVA F(2, 547)=0.008, p=0.992, <0.001.The 0.23, point maximum difference (undergraduate to postgraduate) accounts for 0.3% of the scale range, thus confirming nonneglectibility of education direct effect on CB only after consciousness has been established. This finding is in sharp contrast with the 10.2% income effect, revealing that financial means rather than educational qualifications are the main obstacle to converting health awareness into purchasing behavior. Figure 5 presents demographic moderation figures in the form of violin plots as well as interaction effects. The gender comparison reveals overlapping distributions with females median being slightly higher and males variance wider. Income groups have distinctly separated medians and variances which are visualized through split violins showing density on both sides. Education groups have completely overlapping near, identical distributions, perfectly matching the null effect. To visualize the interaction, the panel plots HFC, CB relationship separately for each income group, showing parallel slopes with vertical displacement, i.e. income acts as an intercept shifter rather than a slope modifier (no interaction).Low, income individuals would need around 13 more HFC points than high, income individuals to be at the same level of CB, i.e. the extent of the financial constraint is being quantified.

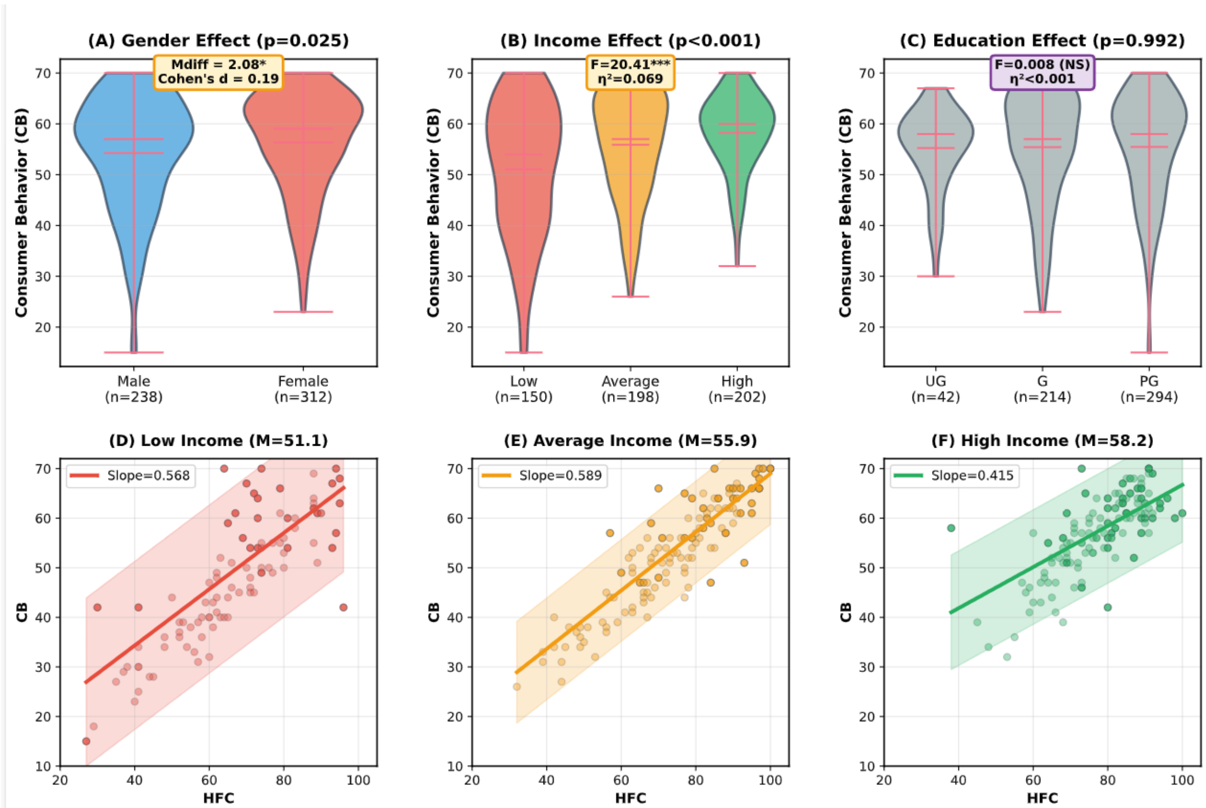


Figure 5. Demographic influences on consumer behavior and income-based HFC–consumer behavior analysis.

Figure 5 presents demographic moderation figures in the form of violin plots as well as interaction effects. The gender comparison reveals overlapping distributions with females median being slightly higher and males variance wider. Income groups have distinctly separated medians and variances which are visualized through split violins showing density on both sides. Education groups have completely overlapping near, identical distributions, perfectly matching the null effect. To visualize the interaction, the panel plots HFC, CB relationship separately for each income group, showing parallel slopes with vertical displacement, i.e. income acts as an intercept shifter rather than a slope modifier (no interaction).Low, income individuals would need around 13 more HFC points than high, income individuals to be at the same level of CB, i.e. the extent of the financial constraint is being quantified.

4.6 Computational Methodology

Python 3.12.1 with scikit, learn 1.3.2, statsmodels 0.14.1, and pandas 2.1.4 libraries were used to perform all the analyses, and Linux Ubuntu 24.04 platform was the operating system of the machine. The entire analytical process took 183 seconds on a computer with Intel Core i7 processor and 16GB of RAM.

The statistical models had 9 predictors (5 psychological scales, 4 demographics) and 550 observations, thus the 61:1 observation, to, predictor ratio was more than adequate. Random Forest used Gini impurity criterion together with a minimum samples split of 2 and bootstrap aggregation. Cross, validation employed a stratified 5, fold partitioning with random state 42 for reproducibility. All hypothesis testing followed two, tailed significance at =0.05 with Bonferroni correction for multiple comparisons where necessary. K, Means clustering adopted Euclidean distance metric along with k, means++ initialization and a maximum of 300 iterations. Convergence was reached after 15 iterations for k=4 solution. Model selection was done on the basis of the performance of the test set so as to avoid over fitting, with the final evaluation on the held, out 20% data which was never used during training or hyper parameter tuning.

Table 2. Summary of Research Objectives and Key Findings

Objective	Relationship / Analysis	Method Used	Key Finding	Evidence Section
O1	HFC → Consumer Behavior	Correlation & Linear Regression	Strong positive relationship	Section 4.1
O2	Advertisement as mediator	Mediation analysis	Weak mediation effect	Section 4.2
O3	Social media as moderator	Moderation regression	Positive moderation effect	Section 4.2
O4	Predictive performance	ML vs Linear Models	Random Forest outperformed others	Section 4.3
O5	Consumer segmentation	K-means clustering	Four distinct consumer segments identified	Section 4.4
O6	Demographic influence	ANOVA / t-tests	Income significant; others negligible	Section 4.5

To make the paper clearer and give a single account of the results, we outline in Table 2 the main research questions, methodological tools, and major results. The table makes it explicit how each objective can be verified by different parts of the study, thus connecting health consciousness, digital influence mechanisms, predictive modeling, consumer segmentation, and demographic effects in a single analytical framework..

5. CONCLUSION

This paper offers a detailed research framework to study the interplay between health and fitness consciousness and consumer behavior by combining statistical and machine learning approaches. The results of this paper support the conclusion that psychological factors especially health consciousness and cognitive, affective, behavioral states have a much more influential role in determining the purchasing decisions related to health of Indian millennials than demographic factors.

By mapping out four different types of customers, Digital Engagers, Wellness Enthusiasts, Traditional Consumers, and Disengaged, it has been shown that millennial consumers are not a homogeneous group and that psychographic segmentation is of great use when it comes to health marketing and the design of interventions. Furthermore, the paper provides evidence that social media is the means that makes the transition of health consciousness to consumption stronger, whereas income is the main element limiting the translation of awareness/education into actual usage.The main strengths of this study are its being the first to measure the relatively small intentionbehavior gap of health, conscious millennials, first to identify the role of social media that assists the adoption of health behavior in health, conscious millennials, and finally to demonstrate that financial accessibility is the still the prevailing obstacle to health, oriented consumption.

However, the authors admit the paper has some limitations, e.g., it is based on an urban Indian sample and survey data. Future research may focus on longitudinal studies to establish causality, look into rural and cross, cultural comparisons, and use objective purchase or transaction, level data. Combining state, of, the, art deep learning models with social network analysis can also lead to improved predictive accuracy and behavioral insights. The paper in general offers valuable directions for data, driven health marketing, digital engagement, and policy making which can lead to healthier and sustainable behaviors of millennials.

Conflicts of Interest

"The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper."

Funding Statement : NA

References

1. Huang, Lingchen, Zongsheng Huang, and Bin Liu. "Interacting with strategic waiting for store brand: online selling format selection." Journal of Retailing and Consumer Services 67 (2022): 102987.
2. Gkinko, Lorentsa, and Amany Elbanna. "Designing trust: The formation of employees' trust in conversational AI in the digital workplace." Journal of Business Research 158 (2023): 113707.
3. Pan, Jialiang, Kun-Shan Wu, Di-Man Huang, and Shu-Wen Sun. "Role of Health-Consciousness on Purchase Intention of Health and Wellness Food: The Serial Mediating Effect of Attitude and Desire." Nutrients 17, no. 5 (2025): 746.
4. Wang, Xuequn, Xiaolin Lin, and Bin Shao. "Artificial intelligence changes the way we work: A close look at innovating with chatbots." Journal of the Association for Information Science and Technology 74, no. 3 (2023): 339-353.
5. Hitcham, Lucy, Hannah Jackson, and Richard JE James. "The relationship between smartphone use and smartphone addiction: An examination of logged and self-reported behavior in a pre-registered, two-wave sample." Computers in Human Behavior 146 (2023): 107822.
6. Wu-Ouyang, Biying. "Are smartphones addictive? Examining the cognitive-behavior model of motivation, leisure boredom, extended self, and fear of missing out on possible smartphone addiction." Telematics and Informatics 71 (2022): 101834.
7. Lin, Chien-Wei, and Kalpesh Desai. "Impact of social norms and joint influences of social connection and self-sufficiency motivations on prosocial budgeting." Psychology & Marketing 39, no. 8 (2022): 1579-1598.
8. DeLiema, Marguerite, Yiting Li, and Gary Mottola. "Correlates of responding to and becoming victimized by fraud: Examining risk factors by scam type." International Journal of Consumer Studies 47, no. 3 (2023): 1042-1059.
9. Li, Youyou, and Baoshun Shan. "Exploring the role of health consciousness and environmental awareness in purchase intentions for green-packaged organic foods: an extended TPB model." Frontiers in Nutrition 12 (2025): 1528016.
10. Li, Shu-Chu Sarrina, Tai-Yee Wu, Huai-Kuai Zeng, and Shih-Yu Lo. "Channels adopted for information seeking during COVID-19: Comparing social media with news media and interpersonal communication in Taiwan." International journal of environmental research and public health 19, no. 15 (2022): 9321.
11. Liang, Haisheng, Zhihui Wu, and Shuangjiao Du. "Study on the impact of environmental awareness, health consciousness, and individual basic conditions on the consumption intention of green furniture." Sustainable Futures 8 (2024): 100245.
12. Al-Dahidi, Sameer, Mohammad Alrbai, Hussein Alahmer, Bilal Rinchi, and Ali Alahmer. "Enhancing solar photovoltaic energy production prediction using diverse machine learning models tuned with the chimp optimization algorithm." Scientific Reports 14, no. 1 (2024): 18583.
13. Deniz, Emre, and Semanur Çökekoğlu Bülbül. "Predicting customer purchase behavior using machine learning models." Information Technology in Economics and Business 1, no. 1 (2024): 1-6.
14. Jakhar, Jatin. "Exploring the Impact of Social Media on Consumer Behavior: Trends and Influences." Journal of Advanced Management Studies 2, no. 1 (2025): 46-56.
15. Dzreke, Simon Suwanzy, and Semefa Dzreke. "Influence of Social Media on Consumer Behavior: Investigating How Social Media Platforms Shape Consumer Purchasing Decisions and Loyalty." Open Journal of Business and Management 13, no. 2 (2025): 1154-1175.