IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155270

Predictive Analysis of Health and Lifestyle Patterns using CRM and Machine Learning

Prathamesh Chambole[1], Dr S.R Gupta[2], Dr R.A Kale[3]

Student, Department of Computer Science and Engineering, PRMIT&R, Amravati, India[1]

Professor, Department of Computer Science and Engineering, PRMIT&R, Amravati, India[2]

Assistant Professor, Department of Computer Science and Engineering, PRMIT&R, Amravati, India[3]

Abstract: This study shows how CRM and machine learning can be used to predict health and living trends in order to help with preventive healthcare. The study uses an open source dataset that has factors about demographics, physical exercise, food, sleep, mental health, and medical background. Data preparation was done to make the data better by dealing with missing values, getting rid of duplicate and incomplete records, lowering the number of errors, and getting the dataset ready for reliable model training. It was possible to turn clinical, physiological, and behavioral traits into useful data for machine learning models by using feature engineering and feature extraction. For figuring out health risks and predicting them, Random Forest, Decision Tree, Logistic Regression, XGBoost, and voting-based classification methods were looked at. With AUC values of 0.68 and 0.64, respectively, the ROC results showed that Random Forest did a little better than Decision Tree. The overall risk classification showed that 43.7 percent of people were low risk, 35.4% were intermediate risk, and 20.9 percent were high risk.

Keywords: Health and lifestyle patterns; Machine learning; Preventive healthcare; Random Forest; Decision Tree; Risk stratification.

I. INTRODUCTION

Predictive analysis has become more useful in preventive healthcare as more data about health and habits become available. Health results are affected by many things, such as a person's background, level of physical exercise, food, sleep, mental health, and medical history. When all of these factors are gathered and looked at together, they can give us useful information about people's health and possible risk levels. Raw health data, on the other hand, often has noise, missing values, duplicate records, and problems because it comes from various sources and user actions. Before using machine learning models, they need to be properly preprocessed, their features extracted, and their technical features. This research is mostly about how to use CRM and machine learning to look at formal health and living data and put people into three groups: low risk, intermediate risk, and high risk. Supporting personalized health promotion, ongoing health tracking, and early discovery of people who may need more preventive care and focused action is the goal.

II. METHODOLOGY

The data set used is an open source data set which has organized data set related to health and lifestyle. It contains information about such things as physical exercise, food, sleep patterns, mental health and medical background. It's aimed at data scientists, experts and people working in machine learning who want to know more about how lifestyle choices impact health as a whole.
Some important features include:

- Demographics: age, gender, BMI and health factors related to the job.
- How often you exercise, the number of steps you take each day, and the amount of time you sit.
- Dietary Habits Water intake Fruits, vegetables and processed foods you eat.
- Sleep Patterns: How long you sleep on average, the quality of your sleep and how often you go to bed.
- Mental health: levels of stress, frequency of meditation, and reported happiness.
- Medical history. Includes common illnesses, medicines and habits for preventative care.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155270

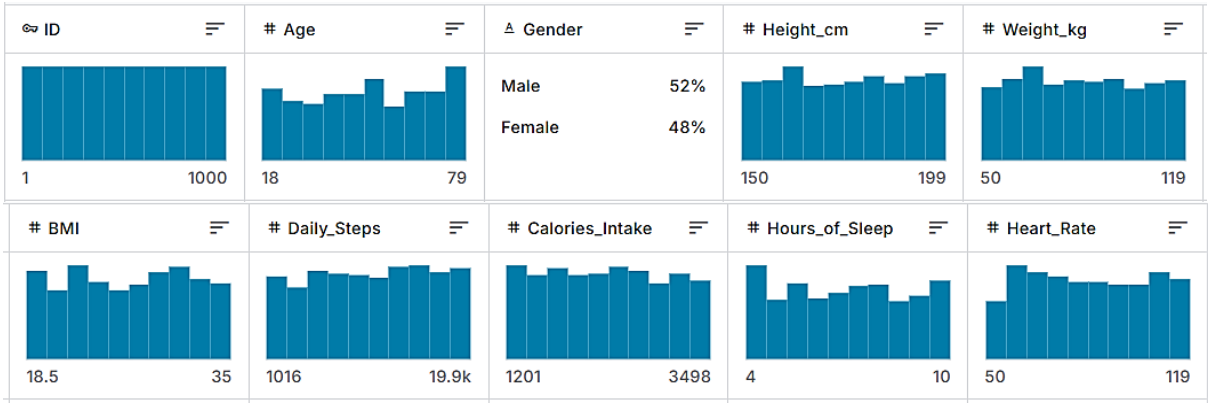

Figure 1: Health and Lifestyle dataset features

The raw data collected are noisy, missing numbers, and inconsistent due to the differences in acquisition sources and user behavior. The preprocessing is therefore aimed at improving the reliability and quality of the data. Data is cleaned of duplicates or missing data and missing data is filled up by appropriate statistical imputation methods. Identifying outliers and removing them improves model performance. Wearable sensor data are synchronized temporally with clinical records so as to be consistent across time-series data. Data extraction is a process that assists in categorizing disease risk and analyzing preventive healthcare by selecting relevant clinical, physiological, and behavioral factors. Neural network models require processed data to be fed through feature engineering from raw health data. Clinical aspects include vital signs, test biomarkers and evidence of chronic disease, which can be extracted from electronic health data. Statistical features gathered from wearable sensors include resting heart rate, heart rate variability, activity intensity, sleep efficiency, and even inactive behavior. A numerical or categorical representation can be used to code a variety of aspects of a person's lifestyle, including stress levels, indicators of dietary quality, physical activity scores, and behavioral risk factors. Feature normalization and scaling is used to scale all features into the same range. Dimensionality reduction approaches and correlation analysis are used to eliminate redundant and superfluous features for faster computation and better generalization of the model.

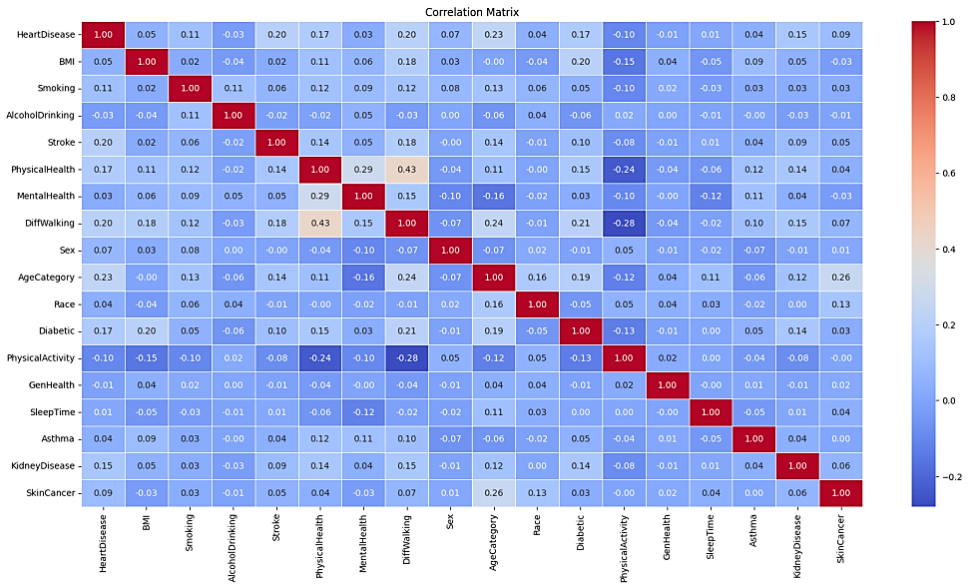

Figure 2: Correlational Matrix for the dataset

The extracted and selected characteristics are then combined into one dataset that represents the health profile of each individual. The information is tagged based on past medical outcomes and professional recommendations to categorize individuals into different health risk categories like low risk, medium risk and severe risk. This is valuable for supervised learning applications in disease risk prediction and personalized health promotion. After this, the dataset is split into training, validation and test subsets in order to have an objective evaluation of the model and to prevent overfitting during training of the model.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155270

In the proposed preventative healthcare system, many models based on machine learning are used to enable individualised health promotion, continuous health monitoring and accurate prediction of the probability of sickness. These models are built to analyze different types of healthcare data, including clinical data, measurements from connected devices, and lifestyle data, by learning complex relationships between various health indicators. An example of an ensemble method is a random forest (RF) which uses a number of different decision trees. A prediction of a new sample x′ can be computed as:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

fb(x') is the bth tree and B is the number of trees. This method combined with GLOVE word embeddings outperformed its competitors on the gender prediction challenge. Decision Trees (DT) with ID3 allow recursive partitioning of the feature space. This allows for interpretable models. The impurity reduction is calculated for a node m as:

$$\Delta i(m) = i(m) - \frac{N_{left}}{N_m} i(m_{left}) - \frac{N_{right}}{N_m} i(m_{right})$$

the number of samples (N), and the impurity measure (i(m)), which might be Gini impurity or entropy, for example. The ensemble approaches used in this work were constructed using decision trees as their foundation. The logistic function is used to represent the likelihood of class membership in logistic regression (LR) for binary tasks:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The parameters of the model are β0 and β1. Our strategy was well substantiated in both the sentiment classification problem and the gender prediction challenge. XGBoost (XGB) is a gradient boosting based implementation with a regularization term:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where fk denotes the k-th tree, l denotes the loss function and Ω denotes the regularisation term. The textclassification tasks in this research showed that XGBoost is very good at handling complex feature interactions. Added to combine the predictions of multiple base classifiers. The voting process may also be characterized as follows:

$$\text{Hard Voting: } y = \text{mode}\{C_1(x), C_2(x), ..., C_n(x)\}$$

$$\text{Soft Voting} : y = \arg\max_i \sum_{j=1}^{n} w_j p_{ij}$$

is the prediction of the i-th classifier, are its weights, are its probability estimates.

The Decision Tree model classifies people into groups using a sequence of rules based on features, as shown in the classification trees. The first split in Figure 3 is a major input variable and then “Gini values, sample counts and class value distributions” are used to split the dataset into smaller branches. The colored nodes represent the model’s risk decision making process, and the leaf nodes represent the final classification results. Since each decision chain can be traced back through the root node to the final class, the model is easy to understand. The depth and many splits indicate the model’s sensitivity to feature variation and potential need for validation to avoid overfitting.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155270

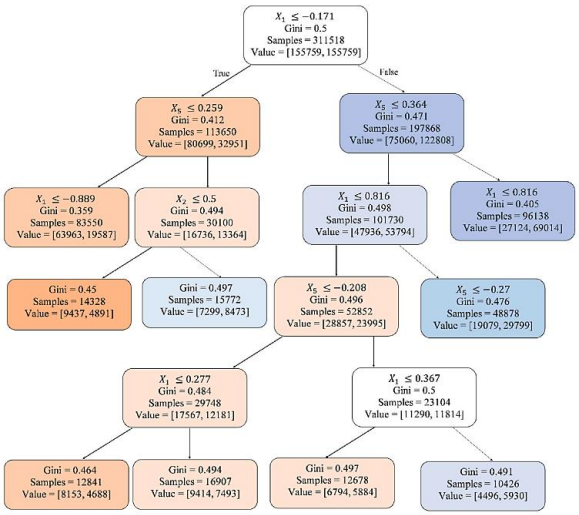

Figure 3: DT dataset classification

III. RESULTS AND DISCUSSION

The ROC comparison shows that both models can tell the difference between health risk classes to a modest degree, but the Random Forest model does a little better than the Decision Tree model. Figure 4 shows that the blue RF curve stays above the orange DT curve over a number of false positive rate intervals. This means that the system is more sensitive at the same error levels. Because RF's area under the curve is 0.68 and DT's is 0.64, it seems that RF gets more useful trends from the health and lifestyle factors that were used in the study. The numbers aren't very high, though, which means the prediction job is still hard and might need more features, better preparation, or model tuning. The finding shows that RF is the more accurate model for identifying preventable health risks in this dataset.

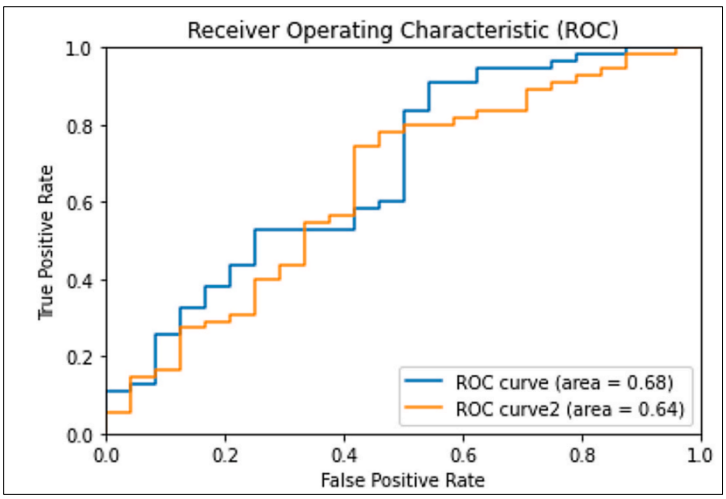

Figure 4: RF (Blue) and DT (Orange) ROC curves

The RF model results show that clinical and social factors have different effects on projection. Figure 5 shows that CH and VT are the most important factors, which means they have a big effect on how the model classifies risk. HB also makes a big difference, but WT, DS, ST, DD, EF, age, and gender have much smaller effects. This shows that the Random Forest model doesn't give equal weight to all traits, but instead relies on a few key predictions. Compared to RF, the DT model shows a more spread-out trend of importance. Figure 6 shows that WT has the most significant effect, followed by EF and CH. VT and age also play a significant role. It's not very important that HB and gender are important. This means that the Decision Tree model sorts things based on a number of factors, but it still relies mostly on certain health and living markers.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155270

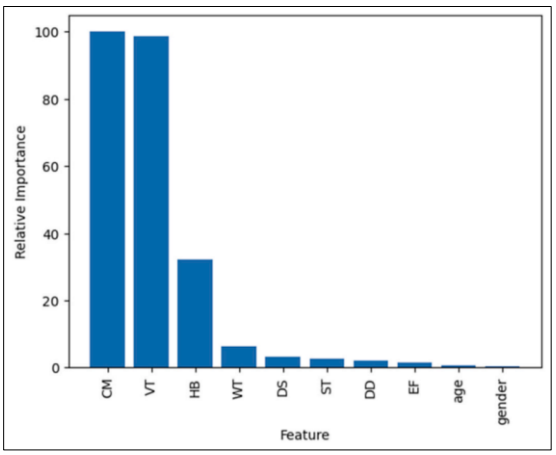

Figure 5: RF Model results

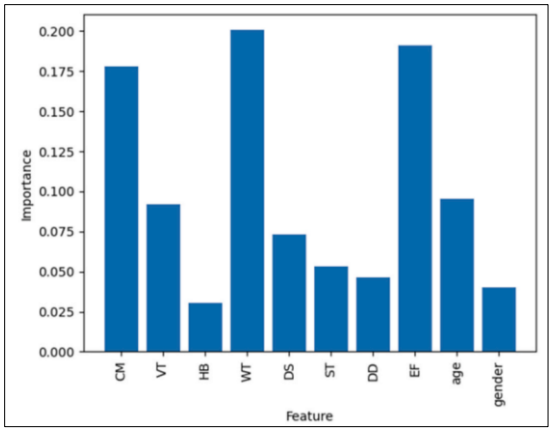

Figure 6: DT model results

The LR model indicates a balanced contribution of factors in the prediction process. ST has the most relevance in Figure 7, followed by EF and VT closely, and then CH, gender, HB, WT, DD, DS and age, albeit with lower variance than the tree based models. That is, Logistic Regression gives a more fair distribution of predictive weight to the features that are available.

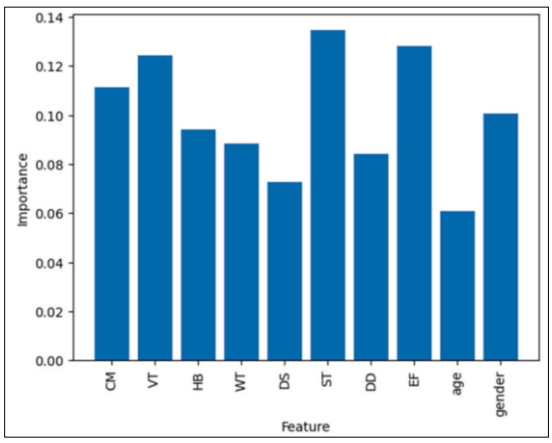

Figure 7: LR model results

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155270

The result of the risk stratification reveals that majority of the people belong to the low risk category, followed by the moderate risk group and lastly the high risk group. Table 1 presents 43.7 percent low risk, 35.4 percent moderate risk, and 20.9 percent high risk. This distribution is visually supported by the pattern in figure 8, with the risk percentage decreasing progressively from low to high danger. This suggests that the dataset has a higher percentage of relatively healthy people but still a significant number is in the intermediate and high risk groups. These findings are valuable for preventative health planning because they assist identify populations who may need more monitoring and focused intervention.

TABLE I RISK STRATIFICATION RESULTS

Risk Level	Percentage (%)
Low Risk	43.7
Moderate Risk	35.4
High Risk	20.9

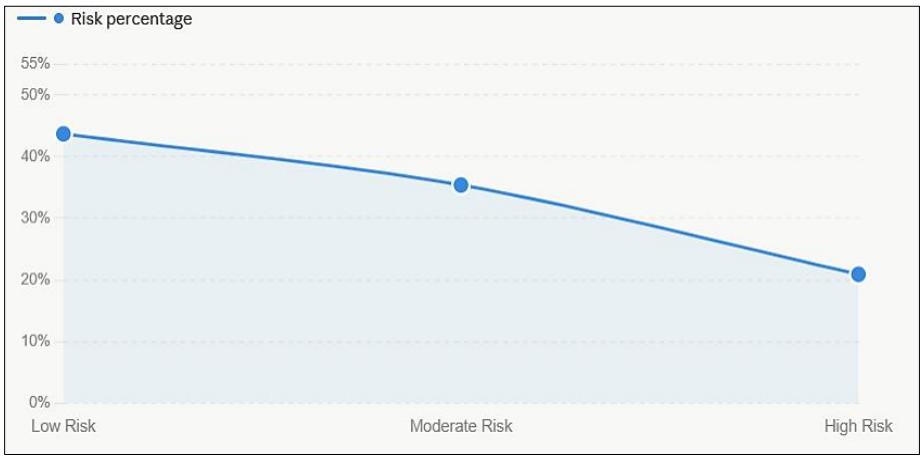

Figure 8: Risk Stratification Results

IV. CONCLUSION AND FUTURE SCOPE

This research shows that machine learning may be successfully used for the analysis of health and lifestyle trends and for aiding preventative healthcare risk categorization. The data set integrated demographic, behavioral, clinical and lifestyle related characteristics to describe each person's health profile. After preprocessing, feature engineering and model creation, the findings indicated that various algorithms picked up health risk patterns differently. The ROC analysis suggested that Random Forest fared better than Decision Tree with AUC of 0.68 against 0.64, demonstrating a superior capacity to discriminate across risk groups. Feature significance findings also demonstrated that the chosen variables had a larger contribution to prediction, notably in Random Forest and Decision Tree models. Logistic Regression spread the prediction weight more equally throughout the available features. The risk stratification findings indicated that the majority of the people were in the low risk category, however the moderate and high risk categories still comprised a considerable proportion of the dataset. These results imply that machine learning based analysis might be useful to identify people who may need monitoring, early assistance and focused preventive healthcare intervention.

REFERENCES

[1]. Unanah, O. V., & Mbanugo, O. J. (2025). Integration of AI into CRM for effective US healthcare and pharmaceutical marketing. *World Journal of Advanced Research and Reviews*, *25*(02), 609-630.

[2]. Islam, M. M., & Shamsuddin, R. (2021). Machine learning to promote health management through lifestyle changes for hypertension patients. *Array*, *12*, 100090.

[3]. Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*, *10*(1), 40.

[4]. Cavalcanti Albuquerque Brayner, L., & Pacheco, E. (2024). Using Predictive Analytics to identify risk of Heart Disease based on lifestyle factors and health metrics.

DOI: 10.17148/IJARCCE.2026.155270

[5]. Valli, L. N. (2024). Predictive analytics applications for risk mitigation across industries; a review. *BULLET: Jurnal Multidisiplin Ilmu*, *3*(4), 542-553.

[6]. Unanah, O. V., & Mbanugo, O. J. (2025). Integration of AI into CRM for Effective U.S. healthcare and pharmaceutical marketing. *World Journal of Advanced Research and Reviews*, *25*(2), 609–630. https://doi.org/10.30574/wjarr.2025.25.2.0396

[7]. Dixon, D., Sattar, H., Moros, N., Kesireddy, S. R., Ahsan, H., Lakkimsetti, M., Fatima, M., Doshi, D., Sadhu, K., & Hassan, M. J. (2024). Unveiling the Influence of AI Predictive analytics on patient Outcomes: A Comprehensive Narrative review. *Cureus*, *16*(5), e59954. https://doi.org/10.7759/cureus.59954

[8]. Kwon, H., Kim, H. H., An, J., Lee, J. H., & Park, Y. R. (2021). Lifelog data-based prediction model of digital health care app customer churn: retrospective observational study. *Journal of medical Internet research*, *23*(1), e22184.

[9]. Dorgbefu, E. A. (2021). Enhancing customer retention using predictive analytics and personalization in digital marketing campaigns. *Int J Sci Res Arch*, *4*(1), 403-23.

[10]. Park, C., Joo, J., You, O., Yi, S., Kim, C., & Jo, A. (2024). Development of a predictive model for managing lifestyle behaviors among patients with chronic skin diseases: Using machine learning techniques. *Informatics in Medicine Unlocked*, *48*, 101528. https://doi.org/10.1016/j.imu.2024.101528

[11]. Segun-Falade, O. D., Osundare, O. S., Kedi, W. E., Okeleke, P. A., Ijomah, T. I., & Abdul-Azeez, O. Y. (2024). Utilizing machine learning algorithms to enhance predictive analytics in customer behavior studies. *International Journal of Scholarly Research in Engineering and Technology*, *4*(1), 001-018.

[12]. Taherkhani, L., Daneshvar, A., Amoozad Khalili, H., & Sanaei, M. (2025). Analysis and Optimization of Customer Lifetime Value Prediction using Machine Learning and Deep Learning Models by RFM Techniques. *International Journal of Web Research*, *8*(2), 79-92.

[13]. Onikoyi, B. Q. (2025). Exploring Predictive Models of Consumer Behaviour Using Machine Learning, NLP, and Data Mining.

[14]. Modi, K., Singh, I., & Kumar, Y. (2023). A comprehensive analysis of artificial intelligence techniques for the prediction and prognosis of lifestyle diseases. *Archives of Computational Methods in Engineering*, *30*(8), 4733–4756. https://doi.org/10.1007/s11831-023-09957-2

[15]. Madanchian, M., Taherdoost, H., & Rafiee, A. (2026). Predictive Analytics for Customer Behavior in AI Insurance. *Transforming Risk*, 87–102. https://doi.org/10.1201/9781003621119-4

[16]. Ashok, P., Sridevi, S. L., Gorli, R., Krishna, K. M., Abinaya, D., & Sidhu, P. (2025). Integrating Advanced Machine Learning Techniques for Enhanced Customer Lifetime Value Prediction. *2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS)*, 1–6. https://doi.org/10.1109/worldsuas66815.2025.11199214

[17]. Droomer, M. (2020, December 1). *Predicting the next purchase date for an individual customer using machine learning*. https://scholar.sun.ac.za/handle/10019.1/109221

