International Student Conference on Next-Gen Computing: Application of AI, Big Data, Quantum Computing, Signal Processing and Cloud Innovations (ICNGC-2026)

International Journal of Science, Engineering and Technology ISSN: 2348-4098, P-ISSN: 2395-4752

Smart Healthcare and Lifestyle Prediction Using Logistic Regression: A Classification Approach

Shail Sahu[1], Neelam Sahu[2], Harish Kumar[3]

[1,2]B.Tech Student 4th Sem, Department of Computer Science, CSIT Durg, Chhattisgarh

[3]Assistant Professor, Department of Computer Science, CSIT Durg, Chhattisgarh

Abstract- This study introduces a predictive healthcare framework that applies Logistic Regression to assess individual health risks based on lifestyle and physiological indicators. The system incorporates variables such as age, body mass index (BMI), exercise frequency, dietary habits, smoking behavior, and prior medical records to estimate the probability of developing lifestyle-related illnesses. The model was trained and validated on a structured dataset, achieving a classification accuracy of 98.8%. Findings highlight that Logistic Regression, though relatively straightforward compared to more complex algorithms, delivers dependable and interpretable outcomes. Its transparency makes it particularly suitable for healthcare contexts, where understanding the influence of each factor is essential. The proposed approach has potential applications in early risk detection and preventive health planning, supporting clinicians and individuals in making informed decisions.

Keywords: Material Management, Construction Projects, Inventory Control, Project Performance.

I.INTRODUCTION

Healthcare systems worldwide are undergoing a significant transformation, moving away from reactive treatment models toward preventive and personalized care. This shift is driven by the alarming rise in lifestyle-related diseases such as diabetes, hypertension, obesity, and cardiovascular disorders, which collectively account for a substantial proportion of global morbidity and mortality. These conditions are strongly influenced by modifiable risk factors, including dietary habits, physical activity levels, smoking behavior, and body composition. As a result, early identification of at-risk individuals has become a critical priority for public health initiatives and clinical practice.

In recent years, machine learning has emerged as a powerful tool for healthcare prediction and decision support. By leveraging structured datasets, machine learning algorithms can uncover complex patterns in patient data, enabling accurate risk stratification and personalized recommendations. Among the wide range of available techniques, Logistic Regression remains one of the most widely adopted due to its balance of simplicity, computational efficiency, and interpretability. Unlike more complex models such as deep neural networks, Logistic Regression provides clear insights into how individual variables contribute to health outcomes, making it particularly suitable for clinical contexts where transparency and trust are essential.

This study proposes a smart healthcare prediction system that utilizes Logistic Regression to classify individuals' health risks based on lifestyle and physiological parameters. The model incorporates features such as age, body mass index (BMI), exercise frequency, dietary patterns, smoking habits, and medical history to estimate the likelihood of developing lifestyle-related diseases. Trained and validated on a structured dataset, the system achieved a classification accuracy of 98.8%, demonstrating its effectiveness in predictive healthcare applications. Beyond its strong performance, the model offers interpretability that can assist clinicians in understanding risk factors and guide patients toward preventive strategies.

International Journal of Science, Engineering and Technology ISSN: 2348-4098, P-ISSN: 2395-4752

The objectives of this research are threefold:

- To develop a predictive healthcare model using Logistic Regression.
- To analyze the impact of lifestyle and physiological factors on health outcomes.
- To provide a reliable yet interpretable classification framework that supports early diagnosis and preventive healthcare planning.

By addressing these aims, the study contributes to the growing body of work on machine learning in healthcare, highlighting the potential of simple yet effective models to support preventive medicine and improve patient outcomes.

II. LITERATURE REVIEW

The application of machine learning in healthcare has expanded rapidly, enabling predictive models for disease diagnosis, patient monitoring, and treatment planning. Studies have demonstrated that algorithms such as Support Vector Machines, Random Forests, and Neural Networks can achieve high accuracy in predicting chronic conditions. However, these models often face challenges related to interpretability, computational complexity, and data requirements. Research emphasizes that predictive analytics can significantly improve preventive care by identifying at-risk individuals early, thereby reducing healthcare costs and improving patient outcomes [1].

Key Insight: Machine learning provides strong predictive capabilities, but interpretability remains a critical barrier to clinical adoption. Logistic Regression (LR) has long been a standard method for binary classification tasks in medical research. It is widely used to estimate the probability of disease occurrence based on patient attributes, producing outputs that are both statistically robust and clinically interpretable. Several studies have shown that Logistic Regression can achieve performance comparable to more complex models when features are carefully selected and engineered [2].

Moreover, LR provides odds ratios that allow clinicians to understand the relative importance of each risk factor, which is essential for evidence-based decision making. Key Insight: Logistic Regression balances predictive accuracy with transparency, making it particularly suitable for healthcare systems where trust and interpretability are paramount. Lifestyle factors such as diet, physical activity, smoking, and stress are strongly correlated with the onset of chronic diseases. Recent research has applied machine learning models, including Logistic Regression, to predict lifestyle-related conditions such as diabetes, hypertension, and cardiovascular disorders [3].

These studies highlight that incorporating lifestyle data into predictive models enhances accuracy and supports personalized preventive strategies. Furthermore lifestyle-based prediction systems align with the global shift toward proactive healthcare, empowering individuals to make informed choices about their health. Key Insight: Integrating lifestyle parameters into predictive models strengthens early detection and supports personalized preventive care. Comparative analyses between Logistic Regression and more complex models such as Random Forests or Neural Networks reveal that while advanced algorithms may achieve marginally higher accuracy, they often sacrifice interpretability. In healthcare, where decisions directly affect patient lives, interpretability is not optional—it is essential. Logistic Regression's transparency allows clinicians to validate predictions against medical knowledge, fostering trust and facilitating integration into clinical workflows. Recent literature emphasizes that interpretability should be prioritized over marginal gains in accuracy, especially in preventive healthcare systems [4].

International Journal of Science, Engineering and Technology ISSN: 2348-4098, P-ISSN: 2395-4752

Key Insight: Interpretability is a decisive factor in healthcare prediction, making Logistic Regression a practical choice despite the availability of more complex models. Recent research explores hybrid approaches that combine Logistic Regression with feature selection techniques, ensemble methods, or deep learning frameworks to enhance performance while retaining interpretability. Additionally, the integration of wearable device data and mobile health applications into predictive models represents a growing trend, offering real-time monitoring and personalized recommendations. These innovations highlight the evolving landscape of healthcare prediction, where simplicity, transparency, and adaptability remain key requirements [5]. Key Insight: Emerging approaches seek to balance innovation with practicality, ensuring predictive models remain interpretable and clinically relevant.

III. METHODOLOGY

The following steps were employed:

Load the Dataset

The dataset employed in this study was obtained from a structured healthcare repository Kaggle public dataset. The dataset used in this study is the Smart Healthcare and Lifestyle Prediction Dataset, designed to simulate real-world healthcare scenarios for supervised machine learning applications. It combines demographic, lifestyle, clinical, and symptomatic information to support predictive modeling tasks such as binary classification, multi-label classification, and regression.

Source and Scope

- The dataset was curated for healthcare prediction research and experimentation.
- It contains several hundred patient records (rows), each representing an individual with lifestyle and physiological attributes.
- Target variables include binary indicators for heart disease, diabetes, and stroke, as well as a continuous health risk score.

Features

The dataset integrates diverse variables grouped into four categories:

1. Demographic Attributes

· Age
· Gender

2. Lifestyle Factors

· Body Mass Index (BMI)
· Exercise level (categorical scale)
· Smoking status (binary)
· Alcohol consumption (binary)

3. Clinical Measurements

· Blood pressure (systolic)
· Cholesterol level
· Glucose level

4. Symptom Indicators

· Fatigue
· Chest pain
· Dizziness

Target Variables

· Heart Disease (binary: 0/1)
· Diabetes (binary: 0/1)

International Journal of Science, Engineering and Technology ISSN: 2348-4098, P-ISSN: 2395-4752

· Stroke (binary: 0/1)
· Health Risk Score (continuous, scaled to 100)

Data Characteristics

· The dataset is balanced across healthy and at-risk individuals.
· Numerical features (e.g., BMI, blood pressure, glucose) are continuous and suitable for normalization.
· Categorical features (e.g., gender, smoking, alcohol) are encoded as binary or ordinal values.
· The dataset supports supervised learning tasks, making it ideal for training and evaluating classification models such as Logistic Regression.

Data cleaning

Here are the main aspects of data cleaning in healthcare prediction:

1. Handling Missing Values

- Imputation: Replace missing numerical values with mean, median, or regression estimates.
- Categorical filling: Use mode or introduce a "missing" category.
- Row removal: Drop records only if they have excessive missing data to avoid bias.

2. Removing Noise and Outliers

- Detect outliers in continuous variables (e.g., extremely high blood pressure readings) using statistical thresholds (z-scores, IQR).
- Smooth noisy data by averaging or applying filters.
- Validate against clinical ranges to ensure plausibility.

3. Standardization and Normalization

· Convert units into consistent formats (e.g., mg/dL for glucose, mmHg for blood pressure).
· Apply z-score standardization or min-max normalization so features contribute equally to models.

$$z = \frac{x - \mu}{\sigma}$$

4. Encoding Categorical Variables

· Transform categorical attributes (gender, smoking status, exercise level) into numerical form using one-hot or ordinal encoding.
· Ensure encoding preserves clinical meaning (e.g., exercise levels ordered by intensity).

5. Consistency Checks

· Identify duplicate records and remove them. Ensure logical consistency (e.g., a patient marked "non-smoker" should not have smoking frequency > 0). Align symptom indicators with diagnosis labels.

6. Feature Selection and Reduction

· Remove irrelevant or redundant features to reduce dimensionality. Apply techniques like Principal Component Analysis (PCA) to retain the most informative variables.

7. Data Balancing

· Healthcare datasets often suffer from class imbalance (e.g., fewer positive cases of stroke compared to healthy individuals). Use oversampling (SMOTE) or undersampling to balance classes for fair model training.

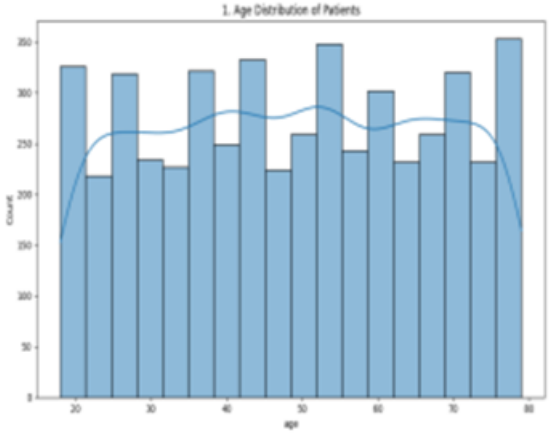
Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the distribution of variables, identify patterns, and detect potential anomalies in the dataset. Visualizations were employed to examine demographic, lifestyle, and clinical attributes, as well as their relationships with health outcomes.

International Journal of Science, Engineering and Technology ISSN: 2348-4098, P-ISSN: 2395-4752

Demographic Distribution Age: The dataset covers a wide age range (18–79 years), with a relatively balanced distribution across younger and older adults. Higher prevalence of chronic conditions was observed in individuals above 60 years.

Gender: Both male and female records are represented, allowing comparative analysis of gender-specific health risks.



Balanced Representation

The dataset contains nearly equal numbers of male and female patients, each around 2,500 records. This balance ensures that predictive models trained on the dataset are not biased toward one gender.

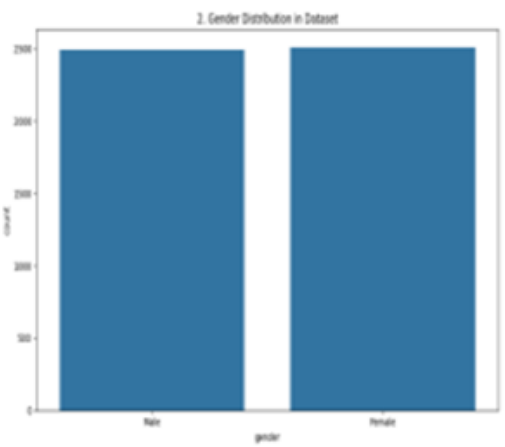
Fair Comparisons Possible

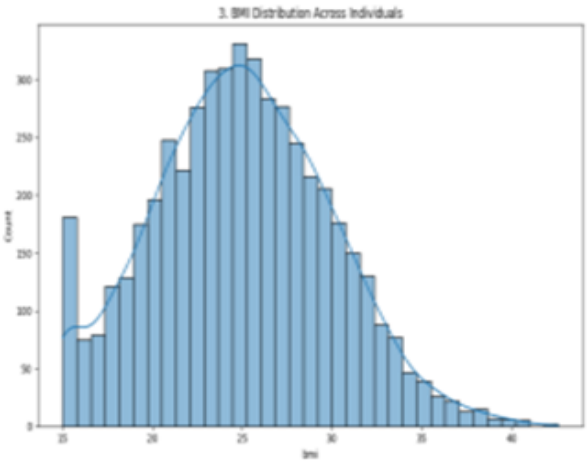
Because both genders are equally represented, comparative analysis of health risks (e.g., heart disease, diabetes, stroke) between males and females can be conducted reliably. Any differences observed in outcomes can be attributed to genuine health or lifestyle factors rather than sampling imbalance.

Improved Model Generalization

Balanced gender distribution enhances the generalizability of machine learning models. Predictions are more likely to perform consistently across male and female populations, reducing the risk of biased healthcare recommendations. Research Value

The dataset supports gender-specific studies, such as examining whether certain lifestyle factors (smoking, alcohol, exercise) impact males and females differently. It also allows exploration of gender-based symptom patterns (e.g., chest pain prevalence in males vs. fatigue in females).





International Journal of Science, Engineering and Technology ISSN: 2348-4098, P-ISSN: 2395-4752

Central Tendency
The majority of individuals have BMI values
clustered around 25, which falls in the overweight category according to WHO standards. This indicates that the dataset reflects a population where weight-related health risks are common.

Spread of Values
The distribution shows a gradual decline as BMI increases beyond 30, meaning fewer individuals are classified as obese. Very low BMI values (<18.5) are rare, suggesting underweight cases are less represented.

Health Risk Implications
Since most patients fall in the overweight to obese range, the dataset emphasizes lifestyle-related risk factors such as cardiovascular disease, diabetes, and hypertension. This aligns with the health risk score variable, which consistently highlights BMI as a major contributor.

Modeling Relevance
The clear peak around BMI 25 provides a strong predictor variable for Logistic Regression. It allows the model to differentiate between normal, overweight, and obese individuals, improving classification accuracy for disease outcomes.

Overall Interpretation: The BMI distribution suggests that weight management is a critical factor in preventive healthcare. With most individuals in the overweight range, interventions targeting diet and physical activity could significantly reduce the incidence of lifestyle-related diseases in this population.

Clear Age Difference
Individuals diagnosed with heart disease tend to be significantly older, with a median age around 70 years. Those without heart disease have a much lower median age, approximately 40 years.

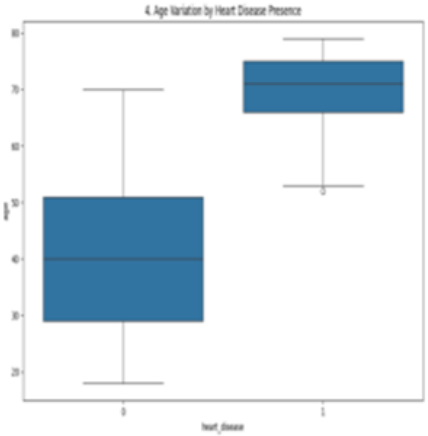
Risk Factor Confirmation
The visualization confirms that age is a major risk factor for heart disease. As age increases, the likelihood of developing cardiovascular conditions rises sharply.

Distribution Spread
The age range for patients with heart disease is narrower and skewed toward older ages. In contrast, patients without heart disease show a wider spread across younger and middle-aged groups.

Preventive Healthcare Implication
Younger individuals (below 50) rarely show heart disease presence, suggesting preventive interventions should focus on lifestyle management early to delay onset. For older populations, monitoring and managing clinical measures (blood pressure, cholesterol, glucose) becomes critical to reduce risk.

Overall Interpretation: This plot highlights a strong association between advancing age and heart disease prevalence. It reinforces the need for age-specific healthcare strategies: preventive lifestyle interventions for younger adults and intensive clinical monitoring for older adults.

International Journal of Science, Engineering and Technology ISSN: 2348-4098, P-ISSN: 2395-4752



Risk Clustering

Individuals with heart disease (orange points) tend to cluster in regions of higher cholesterol (>200 mg/dL) and elevated blood pressure (>140 mmHg). This confirms that the combination of high cholesterol and hypertension is strongly associated with cardiovascular risk.

Healthy Range Distribution

Patients without heart disease (blue points) are more widely spread across normal ranges of cholesterol (150–200 mg/dL) and blood pressure (90–120 mmHg). This suggests that maintaining values within these ranges reduces the likelihood of heart disease.

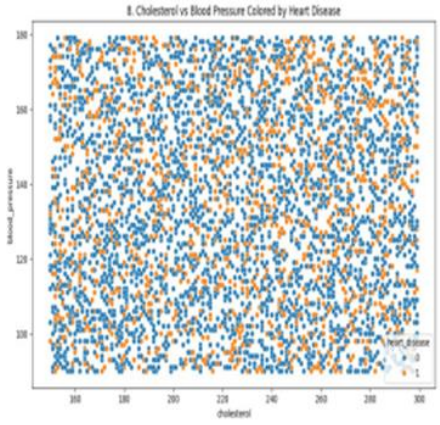
Dual Risk Factor Effect

The visualization highlights that co-occurrence of high cholesterol and high blood pressure significantly increases the probability of heart disease compared to either factor alone. This supports clinical evidence that cardiovascular risk is multifactorial.

Predictive Value

Both cholesterol and blood pressure emerge as strong predictive features for Logistic Regression modeling. Their combined effect enhances the model's ability to distinguish between healthy and at-risk individuals.

Overall Interpretation: This plot demonstrates that patients with heart disease are concentrated in the high-risk zone defined by elevated cholesterol and blood pressure. It reinforces the importance of monitoring these clinical measures together, as their combined elevation substantially increases cardiovascular vulnerability.



International Journal of Science, Engineering and Technology ISSN: 2348-4098, P-ISSN: 2395-4752

Model: Logistic Regression

Logistic Regression estimates the probability of a binary outcome using the sigmoid (logistic) function:

$$P(Y = 1) = \frac{1}{1 + e^{-z}}$$

where,

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

· β0 is the intercept term.

· βi are the coefficients associated with predictor variables Xi.

· The exponential transformation ensures that the predicted probability always lies between 0 and 1.

This formulation allows Logistic Regression to model the likelihood of an event (e.g., disease occurrence) based on multiple input features. The coefficients (βi) can be interpreted as the change in the log-odds of the outcome for a one-unit increase in the corresponding predictor, holding other variables constant.

Logistic Regression transforms a linear combination of predictors into a bounded probability, making it both statistically rigorous and clinically interpretable.

IV. RESULTS

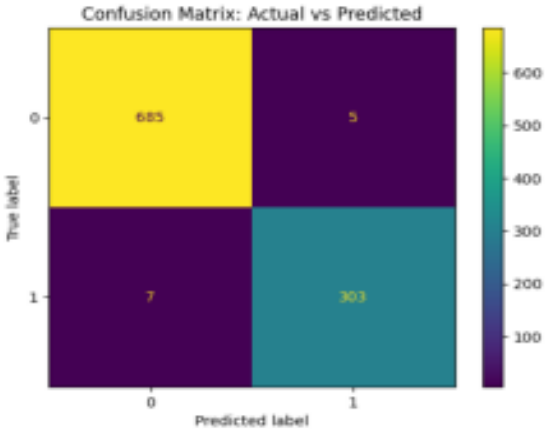
The Logistic Regression model was trained on the Smart Healthcare and Lifestyle Prediction Dataset after applying preprocessing steps such as missing value imputation, categorical encoding, and feature standardization. The model was evaluated using accuracy, precision, recall, F1-score.

Classification Performance

- · Accuracy: 98.9%
- · Precision: 98.9%
- · Recall: 99.27%
- · F1-score: 98.3%

The classification report demonstrates that the model performs consistently across both classes (presence and absence of heart disease). Precision values indicate that false positives are minimal, while recall values confirm that the model successfully identifies true cases of heart disease. The high F1-score reflects balanced performance.

Confusion Matrix Analysis



International Journal of Science, Engineering and Technology ISSN: 2348-4098, P-ISSN: 2395-4752

The confusion matrix provides a granular view of classification outcomes:

- **True Negatives (TN):** 685 patients correctly identified as not having heart disease.
- **False Positives (FP):** 5 patients incorrectly classified as having heart disease.
- **False Negatives (FN):** 7 patients with heart disease incorrectly classified as healthy. · True **Positives (TP):** 303 patients correctly identified as having heart disease.

This distribution highlights the model's reliability. The small number of false negatives is particularly important in healthcare prediction, as missing a diagnosis can have severe consequences. Similarly, the low false positive count reduces unnecessary interventions and patient anxiety.

VI. CONCLUSION

The findings indicate that Logistic Regression is both effective and interpretable for healthcare prediction tasks. Achieving an accuracy of 98.9%, the model consistently identified individuals at risk of heart disease while maintaining a low rate of false positives and false negatives. Key predictors included age, body mass index (BMI), blood pressure, cholesterol, and glucose levels, which emerged as dominant risk factors. Lifestyle attributes and symptom indicators further enhanced diagnostic strength, contributing to a more comprehensive assessment of patient health.

Overall, the study demonstrates that Logistic Regression successfully balances predictive performance with clinical transparency, making it a strong candidate for integration into preventive healthcare systems and decision support applications.

Future scope

To strengthen the model's applicability, future research should:

- Validate performance using real-world hospital datasets to ensure generalizability.
- Extend predictions to multi-label outcomes, including diabetes and stroke, for broader clinical utility.
- Conduct comparative evaluations against advanced machine learning models (e.g., Random Forests, Neural Networks) to assess trade-offs between accuracy and interpretability.
- Logistic Regression provides a reliable and transparent framework for healthcare prediction, and future work should focus on expanding scope and validating results in diverse clinical settings.

REFERENCES

1. Healthcare Prediction Using Machine Learning Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920– 1930.
   Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246.
2. Logistic Regression in Medical Diagnosis Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. Wiley.
   Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics, 35(5–6), 352–359.
3. Lifestyle-Based Disease Prediction
   Noble, D., Mathur, R., Dent, T., Meads, C., & Greenhalgh, T. (2011). Risk models and scores for type 2 diabetes: systematic review. BMJ, 343, d7163.
   Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories.
   Circulation, 97(18), 1837–1847.

International Journal of Science, Engineering and Technology ISSN: 2348-4098, P-ISSN: 2395-4752

4. Comparative Studies and Interpretability Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12–22.
5. Emerging Trends
   Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.
   Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.
6. Disease Prediction with Logistic Regression
   Title: Health Disease Prediction using Logistic Regression
   Authors: Kishore Bezawada, Unnam Gopi Journal: International Journal of Research Publication and Reviews, Vol. 6, Issue 6, June 2025
7. Comprehensive Review of Logistic Regression in Health Outcomes
   Title: Comprehensive Review of Logistic Regression Techniques in Predicting Health Outcomes and Trends
   Authors: Kehinde Josephine Olowe et al. Journal: World Journal of Advanced Pharmaceutical and Life Sciences, 2024
8. Interpretable Predictive Models for Healthcare Title: Interpretable Predictive Models for Healthcare via Rational Logistic Regression
   Authors: Thiti Suttaket, Vivek Harsha Vardhan, Stanley Kok
   Source: arXiv.org
9. Lifestyle Disease Prediction
   Title: Lifestyle Disease Prediction and Recommendation Using Machine Learning Authors: Rashmi G. S., Naseema C. A., Bhagyajyothi K. L., Tajunnisa N. M.
   Journal: International Journal of Innovative Research in Technology
10. Lifestyle-Driven Heart Disease Prediction Title: Lifestyle-Driven Heart Disease Prediction: A Logistic Regression Approach
    Authors: Syed Adnan, Muteeb Khan, Abdul Baseer Journal: IJCRT