Laguna State Polytechnic University

Main Campus

VitaSenseAI: A Machine Learning Model for Lifestyle and Health Risk Prediction

Collaborative Final Project

CSST101 – Machine Learning

CSST102 – Knowledge Representation and Reasoning

Submitted by:

Algobunnies

Kristel N. Ramos

Johann Francois P. Tanyag

Althea Maxene B. Viar

Submitted to:

Mr. Mark P. Bernardino

January 2, 2026

PROJECT OVERVIEW

Lifestyle-related health risks remain a persistent challenge in the Philippines, where factors such as inadequate sleep, smoking, and alcohol consumption significantly contribute to the development of chronic illnesses. These risks are further intensified by limited access to healthcare services, which has reinforced a culture in which many individuals neglect routine medical checkups. As a result, health conditions often remain undetected until they progress into more severe and potentially life-threatening complications. This situation highlights the need for accessible and preventive health monitoring solutions that can support early risk awareness.

To address this concern, the researchers developed VitaSenseAI, a model designed to assess lifestyle-related health risks through the integration of Machine Learning (ML) and Knowledge Representation and Reasoning (KRR). The system utilizes a Random Forest classifier trained on a synthetic lifestyle dataset obtained from Kaggle. The dataset includes key variables such as age, weight, height, body mass index (BMI), exercise frequency, sleep duration, sugar intake, smoking habits, alcohol consumption, marital status, and profession. These features were selected due to their established relevance to health outcomes and lifestyle-related risks.

By combining data-driven predictions with rule-based reasoning, the integration of ML and KRR enhanced the interpretability and reliability of the model's predictions. This hybrid approach allows the system to not only classify health risk levels but also provide logical explanations grounded in health-related rules. The findings of the study demonstrate that hybrid AI systems can play a vital role in preventive healthcare, particularly in resource-constrained settings. Moreover, the project supports Sustainable Development Goal

3 (Good Health and Well-Being) by promoting early health awareness, improving health literacy, and offering accessible wellness assessment tools for local communities.

Keywords: Machine Learning, Knowledge-Based Systems, Lifestyle Analytics, Preventive Health, SDG 3, Random Forest, Expert Systems

OBJECTIVES

General Objective

The study aims to develop a classification model capable of predicting an individual's health risk through the integration of Machine Learning (ML) and Knowledge Representation and Reasoning (KRR). By analyzing user-provided lifestyle and health-related data, the model categorizes individuals into health risk levels, specifically identifying whether they are At-Risk or Not-At-Risk. The use of the Lifestyle and Health Risk Prediction dataset from Kaggle enables the model to generate systematic and reliable risk interpretations. Through this approach, the research promotes proactive health monitoring while providing free and remote access to basic health risk assessment.

Specific Objectives

Encourage and motivate individuals to monitor their health

The research aims to develop a system that addresses lifestyle-related health risks and promotes a preventive health culture, encouraging individuals to regularly monitor and regulate their health through routine self-assessments.

Develop a machine learning model that assesses and interprets lifestyle-related health risks

By integrating Machine Learning with Knowledge Representation and Reasoning, the project

seeks to create a model capable of evaluating individual health data and classifying risk levels into categories such as Low, Medium, High, At-Risk, or Not-At-Risk, while maintaining interpretability through rule-based reasoning.

Provide remote and accessible health risk assessment

Individuals in urban communities and those from low-income households often experience limited access to healthcare services. To address this concern, the researchers developed VitaSenseAI as a free and accessible platform that enables users to perform basic health risk assessments remotely, supporting early awareness without the need for immediate hospital visits.

SYSTEM ARCHITECTURE

The VitaSenseAI system is designed using a hybrid architecture that integrates Machine Learning (ML) and Knowledge Representation and Reasoning (KRR) to deliver accurate, interpretable, and reliable health risk assessments. The overall workflow of the system follows a sequential pipeline, as illustrated below:

User Input → Data Preprocessing → Random Forest Model → Hybrid Decision Logic (KRR Rules) → Final Health Risk Level → Recommendations

User Input

The system begins with the acquisition of user-provided data through a web-based interface. Users are required to input relevant lifestyle, demographic, and basic health information, including age, body mass index (BMI), sleep duration, exercise frequency, smoking status, and alcohol consumption. These variables were selected due to their established relationship with lifestyle-related health risks and their relevance in preventive health assessment.

Data Preprocessing

Prior to model inference, the collected data undergoes preprocessing to ensure data quality and compatibility with the trained machine learning model. This phase includes handling missing or incomplete values, encoding categorical attributes such as smoking and alcohol intake, and normalizing numerical features including age and BMI. Data preprocessing is a critical step that enhances model performance and ensures consistent and reliable predictions.

Random Forest Model

After preprocessing, the refined data is passed to the Machine Learning component of the system, which employs a Random Forest classifier trained on the Kaggle "Lifestyle and Health Risk Prediction" dataset. The model analyzes complex relationships among lifestyle and health indicators to generate an initial health risk classification. This data-driven approach enables the system to capture non-linear patterns and interactions among features that may not be evident through rule-based methods alone.

Hybrid Decision Logic (KRR Rules)

To address the limitations of purely statistical models, VitaSenseAI incorporates a Knowledge Representation and Reasoning layer. This component applies predefined health rules derived from established medical and lifestyle guidelines, such as elevated BMI thresholds, smoking behavior, physical inactivity, and inadequate sleep duration. In cases where the machine learning output conflicts with a critical rule-based condition, the system prioritizes the KRR rules to ensure safety, clinical relevance, and interpretability of the final decision.

Final Health Risk Level

The outputs from the Random Forest model and the KRR reasoning layer are integrated to produce the final health risk classification. This classification categorizes individuals into defined risk levels, such as Low, Moderate, or High Risk, or At-Risk and Not-At-Risk. The

hybrid decision-making process enhances the robustness of the system by combining predictive accuracy with logical justification.

Recommendations

Based on the final health risk level, the system generates rule-based lifestyle recommendations. These recommendations provide users with clear explanations for their risk classification and suggest basic preventive actions, such as improving sleep habits, increasing physical activity, or reducing smoking and alcohol intake. This feature supports early health awareness and encourages proactive health management.

MACHINE LEARNING COMPONENT

Algorithm Used: The system uses Logistic Regression to classify individuals into health risk categories based on lifestyle, demographic, socioeconomic, and clinical features.

Dataset Size: The dataset utilized in this study is obtained from the Kaggle repository titled "Lifestyle and Health Risk Prediction." It consists of approximately 5,000 synthetic records representing individuals with varying lifestyle behaviors, demographic attributes, and basic health indicators.

Model Accuracy: Model performance is evaluated using standard classification metrics, including accuracy, confusion matrix, and cross-validation techniques. These evaluation methods are applied to assess the reliability and generalization capability of the trained model on unseen data.

MACHINE LEARNING PIPELINE

Data Collection: The data for this project is sourced from the publicly available Kaggle dataset Lifestyle and Health Risk Prediction. The dataset contains structured, synthetic data

designed to simulate real-world health and lifestyle scenarios, making it suitable for supervised learning experiments.

Data Preprocessing: Prior to model training, the dataset undergoes preprocessing procedures to improve data quality. These include handling missing values, encoding categorical variables such as smoking status and alcohol intake, normalizing numerical features like BMI and age, and partitioning the dataset into training and testing subsets.

Model Training: The processed data is used to train a Logistic Regression classifier. The model learns the relationship between input features including lifestyle habits, demographic factors, and health indicators and the target health risk level.

Model Evaluation: The trained model is evaluated using quantitative performance metrics such as accuracy scores, confusion matrices, and cross-validation results. These methods help determine the model's predictive capability and identify potential issues such as overfitting.

Model Deployment: After evaluation, the finalized model is deployed using a Flask-based web application. This deployment enables users to input personal health and lifestyle information and receive a predicted health risk level, supporting remote health risk assessment.

DATASET DESCRIPTION

The dataset utilized in this study is a synthetic dataset developed to model realistic health and lifestyle conditions without exposing sensitive personal or medical information.

Dataset Type: A synthetic dataset found in Kaggle was utilized in the project to simulate health-related data for analysis and prediction purposes.

Number of Records: The dataset contains 5,000 records, providing a sufficiently large sample for training and evaluating the machine learning model.

Target Variable: The target variable of the dataset is the Health Risk Category, which classifies individuals into categories such as Low Risk, Moderate Risk, and High Risk based on their health features.

KNOWLEDGE REPRESENTATION & REASONING

This machine learning integrates a rule-based knowledge representation and reasoning components. The rule based system is designed to enhance interpretability and provide logical explanations aligned with established health guidelines.

Rule 1: IF an individual's Body Mass Index (BMI) is greater than 30, THEN the Health Risk Level is classified as High.

Rule 2: IF Smoking Status is Yes and Exercise Frequency is Low, THEN the Health Risk Level is classified as High.

Rule 3: IF Alcohol Intake is Yes and Sleep Duration is less than 6 hours, THEN the Lifestyle Risk is classified as High.

Rule 4: IF Physical Activity is Low and BMI is greater than or equal to 25, THEN the Health Risk Level is classified as Medium to High.

Rule 5: IF Smoking Status is No, Regular Exercise is Yes, and Sleep Duration is at least hours, THEN the Health Risk Level is classified as Low.

HYBRID DECISION LOGIC

The VitaSenseAI system employs a hybrid architecture that merges the predictive power of Machine Learning (ML) with the logical transparency of Knowledge Representation and Reasoning (KRR). This dual-layered approach ensures that the system is

not a "black box" but rather a tool that provides both statistical predictions and human-understandable justifications.

The Decision Workflow

Data Acquisition: The user inputs lifestyle variables (Age, BMI, Smoking Status, Exercise, etc.) via the web interface.

ML Classification Layer: The Logistic Regression (or Random Forest) model processes the inputs to generate a statistical probability of risk (e.g., "At-Risk").

KRR Reasoning Layer: The system passes the same inputs through a predefined Inference Engine containing the health rules (e.g., IF BMI > 30, THEN High Risk).

Conflict Resolution & Integration:

- If the ML prediction and KRR rule align, the system outputs the risk level with high confidence.
- If the ML prediction is "Low" but a KRR rule (like Rule 1 regarding BMI) triggers a "High" classification, the system prioritizes the rule-based logic to ensure safety and clinical relevance.

Output Generation: The final result displays the Risk Category alongside a Rule-Based Recommendation to explain why that risk level was assigned.

SYSTEM FEATURES

VitaSenseAI utilizes a web-based platform to provide users with seamless and remote access to free health risk assessment services. Through the website, individuals with an internet connection can input their lifestyle and health-related data and receive immediate results based on the system’s analysis. The platform processes user inputs using a trained machine learning model enhanced by rule-based reasoning, allowing the system to present both predictions and interpretable insights. The model categorizes the user’s health risk level and displays results in a clear and user-friendly manner.

Key Features

Lifestyle Risk Prediction

The system analyzes user-provided lifestyle data and predicts the individual's health risk level using a supervised machine learning classification model integrated with Knowledge Representation and Reasoning.

Web-Based Platform

VitaSenseAI is accessible through a web browser, enabling users to perform health risk assessments remotely without the need for physical medical checkups.

User Inputs

Users can input relevant personal and lifestyle information such as age, BMI, sleep duration, exercise frequency, smoking status, and alcohol consumption, which are essential in generating accurate risk predictions.

Rule-Based Recommendations

In addition to machine learning predictions, the system applies predefined health rules to explain risk levels and provide basic lifestyle-related recommendations, improving transparency and interpretability.

Visual Studio Code Deployment

The machine learning model is developed and executed using Visual Studio Code, allowing efficient model training, testing, and validation in a cloud-based environment.

TESTING AND EVALUATION

The system was evaluated using a combination of quantitative metrics for the ML component and qualitative verification for the KRR component. Below is a sample of the testing framework used to validate the hybrid logic.

Test Case	Input Summary	Expected Outcome
TC-01: Valid Input – Low Risk	Young adult, normal BMI, non-smoker, adequate sleep, high exercise	Low Risk classification with high confidence
TC-02: Valid Input – Moderate Risk	Middle-aged user, slightly high BMI, occasional exercise, average sleep	Moderate Risk classification
TC-03: Valid Input – High Risk	Older adult, obese BMI, smoker, low exercise, poor sleep	High Risk classification
TC-04: Boundary BMI Case	Normal inputs but BMI near overweight threshold	Moderate Risk due to KRR rule adjustment
TC-05: Smoking Override Rule	Good lifestyle but smoking = “Yes”	Risk level increased by KRR refinement
TC-06: Insufficient Sleep Rule	Low sleep hours (<5) with otherwise moderate inputs	Risk upgraded to Moderate/High Risk
TC-07: Missing Required	One or more required input	Error message indicating

Field	fields missing	missing field
TC-08: Invalid Categorical Input	Profession or lifestyle value not in trained encoder	Validation error returned
TC-09: Extreme Numeric Value	Unrealistically high weight or height	Model handles input or returns safe error
TC-10: Model Confidence Check	Valid input with clear class separation	Prediction returned with confidence level (%)

Table 01. Test Cases

CONCLUSION

The VitaSenseAI project successfully demonstrates the development of a hybrid health risk assessment system that integrates Machine Learning (ML) with Knowledge Representation and Reasoning (KRR) to address lifestyle-related health risks in the Philippines. By leveraging a Random Forest (and Logistic Regression) classifier trained on a synthetic lifestyle dataset and combining it with rule-based reasoning, the system is capable of producing reliable, interpretable, and actionable health risk predictions. The hybrid approach ensures that statistical predictions are complemented by logical, guideline-driven explanations, enhancing user trust and system transparency.

Through its web-based platform, VitaSenseAI provides accessible, remote health risk assessments that encourage proactive monitoring and preventive care, particularly for communities with limited access to healthcare services. The system not only classifies individuals into risk categories such as Low, Moderate, High, At-Risk, or Not-At-Risk but

also delivers clear, rule-based lifestyle recommendations, promoting informed decision-making and early health awareness.

The project highlights the potential of hybrid AI systems in preventive healthcare, supports the achievement of Sustainable Development Goal 3 (Good Health and Well-Being), and underscores the importance of accessible, data-driven tools for improving public health outcomes.

GROUP CONTRIBUTION

Contribution	Member Name
Project Documentation	Ramos, Kristel Tanyag, Johann Francois Viar, Althea Maxene
Literature Review & SDG Alignment	Ramos, Kristel Tanyag, Johann Francois Viar, Althea Maxene
Dataset Analysis & Feature Engineering	Ramos, Kristel Tanyag, Johann Francois Viar, Althea Maxene
Machine Learning Model Development & Training	Ramos, Kristel Viar, Althea Maxene
Knowledge Representation & Rule-Based	Ramos, Kristel

System Design	Viar, Althea Maxene
Back-End Development & System Integration	Ramos, Kristel Viar, Althea Maxene
Front-End Development & User Interface Design	Tanyag, Johann Francois
Web Deployment & Environment Configuration	Tanyag, Johann Francois Ramos, Kristel
Model Evaluation, Testing & Refinement	Ramos, Kristel Viar, Althea Maxene
Logo Design & Visual Identity	Tanyag, Johann Francois
Presentation & Project Defense Materials	Ramos, Kristel Tanyag, Johann Francois Viar, Althea Maxene

Table 02. Member Contribution

REFERENCES

- **Kaggle dataset**

  MIADUL. (n.d.). *Lifestyle and Health Risk Prediction* [Data set]. Kaggle. https://www.kaggle.com/datasets/miadul/lifestyle-and-health-risk-prediction

- **Capstone-Intel article**

  Capstone-Intel Corporation. (2023, September 4). *New Study Reveals Only 40% of Filipinos Get Yearly Checkups, an Issue That Can Be Addressed by UHC Law* [Web page]. Capstone-Intel. https://www.capstone-intel.com/new-study-reveals-only-40-of-filipinos-get-yearly-checkups-an-issue-that-can-be-addressed-by-uhc-law-atty-conti Capstone Intel

- **Inquirer Globalnation article**

  Inquirer Global Nation. (2024, April 8). *WHO: Basic health services lacking in PH, Western Pacific*. Philippine Daily Inquirer. https://globalnation.inquirer.net/230863/who-basic-health-services-lacking-in-ph-western-pacific globalnation.inquirer.net

- **Better Health Victoria page**

  Victorian Department of Health. (n.d.). *Health checks*. Better Health Channel. https://www.betterhealth.vic.gov.au/healthyliving/health-checks