****

**Assessment Report**

on

**“Health Risk Classification”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Name : Anurag

Roll Number : 202401100300057

Section: A

**Under the supervision of**

“BIKKI KUMAR”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

**1. Introduction**

In recent years, lifestyle diseases such as obesity, diabetes, and cardiovascular conditions have become more prevalent due to poor eating habits, irregular exercise, and increasing stress. Accurate prediction of an individual’s health risk based on their lifestyle can lead to early interventions and better health outcomes. This project focuses on building a classification model that categorizes individuals into **low**, **medium**, or **high** risk levels based on three primary features: **BMI (Body Mass Index)**, **exercise frequency**, and **eating habits**.

**2. Problem Statement**

The goal of this project is to **predict the health risk category** of individuals—**Low**, **Medium**, or **High**—based on their BMI, exercise frequency, and eating habits.  
The primary objective is to build a reliable classifier that can help in:

* Early identification of at-risk individuals
* Promoting preventive healthcare
* Supporting personalized health recommendations

**3. Methodology**

1. **Data Collection & Simulation**  
   Since real-world data may not be available, a synthetic dataset was generated consisting of BMI values, weekly exercise frequency, and categorized eating habits.
2. **Data Preprocessing**
   * Encoding categorical variables such as eating habits and risk category into numerical format.
   * Normalization and splitting the dataset into training and testing sets (80:20 ratio).
3. **Model Building**
   * A Random Forest Classifier was chosen due to its robustness and ability to handle non-linear data.
   * The model was trained on the training set and tested on unseen data.
4. **Evaluation**
   * Model performance was assessed using metrics like **Accuracy**, **Precision**, and **Recall**.
   * A **confusion matrix** heatmap was generated to visualize the model's performance across categories.
5. **Visualization**
   * Heat maps and metric reports were used to communicate results.
   * Interpretation of results was done to understand the strengths and weaknesses of the model.

**4. Data Preprocessing**

Due to the unavailability of a publicly accessible dataset, a **synthetic dataset** was created using Python. The dataset includes:

* **BMI**: Continuous numerical values based on a normal distribution.
* **Exercise Frequency**: Discrete integers representing how many times per week a person exercises (0 to 5).
* **Eating Habits**: Categorical values — poor, average, or good.
* **Risk Category**: Target class — low, medium, or high.

**5. Model Implementation**

**Choice of Model**

We selected the **Random Forest Classifier** due to the following advantages:

* It handles both numerical and categorical data efficiently.
* It reduces overfitting by aggregating the predictions of multiple decision trees.
* It is robust to outliers and noise in the data.

**6.Evaluation Metrics**

Once the classification model is trained, it's essential to assess its performance using appropriate evaluation metrics. Since this is a **multi-class classification problem**, we use metrics that give insights into overall accuracy and class-specific performance.

**7. Results and Analysis**

* The model provided reasonable performance on the test set.
* Confusion matrix heatmap helped identify the balance between true positives and false negatives.
* Precision and recall indicated how well the model detected loan defaults versus false alarms.

**8. Conclusion**

The Random Forest classifier demonstrated promising accuracy and consistency in classifying individuals into health risk categories. With further data tuning and real-world health data, its performance can be enhanced for deployment in healthcare or wellness platforms.

**9. References**

* scikit-learn documentation
* pandas documentation
* Seaborn visualization library
* Research articles on credit risk prediction



