

**Project Title:** **Health Risk Classification**   
**Student Name:** Akul Sharma  
**Roll No:** 202401100300027  
**Faculty Supervisor:** Bikki Gupta  
**Institution:** KIET, CSE AI

**1. Introduction**

This report presents an analysis of health risk classification based on Body Mass Index (BMI), exercise hours, and junk food consumption frequency. The goal is to predict whether an individual's health risk level is **low**, **medium**, or **high** using machine learning techniques.

**2. Dataset Overview**

The dataset contains **100 entries** with the following features:

* **BMI** (Body Mass Index)
* **Exercise Hours** (weekly exercise duration)
* **Junk Food Frequency** (weekly consumption frequency)
* **Risk Level** (Target variable: low, medium, high)

**Class Distribution**

| **Risk Level** | **Count** |
| --- | --- |
| **Medium** | 44 |
| **High** | 32 |
| **Low** | 24 |

**Feature Distributions**

* **BMI**: Ranges from **15.3 to 34.8** (some extreme values present).
* **Exercise Hours**: Most people exercise **0-6 hours/week**.
* **Junk Food Frequency**: Varies from **0 (never) to 6 (frequent)**.

**3. Methodology**

**Data Preprocessing**

* **Target Variable Encoding**:
  + low → 0
  + medium → 1
  + high → 2
* **Train-Test Split**:
  + **70% training**, **30% testing**.
* **Feature Scaling**:
  + Standardized using StandardScaler().

**Model Selection**

* **Algorithm**: **Random Forest Classifier** (chosen for robustness in handling non-linear relationships).
* **Evaluation Metrics**:
  + **Accuracy**, **Precision**, **Recall**, **F1-Score**.
  + **Confusion Matrix Heatmap** for visualization.

**4. Results**

**Classification Report**

| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| **Low** | 0.83 | 0.83 | 0.83 | 12 |
| **Medium** | 0.69 | 0.69 | 0.69 | 16 |
| **High** | 0.77 | 0.77 | 0.77 | 13 |

**Overall Metrics**

* **Accuracy**: **75.61%**
* **Precision**: **75.61%**
* **Recall**: **75.61%**
* **F1-Score**: **75.61%**

**Confusion Matrix Heatmap**

  
*(Example visualization showing correct vs. misclassified predictions.)*

**Feature Importance**

| **Feature** | **Importance** |
| --- | --- |
| **BMI** | 0.65 |
| **Exercise Hours** | 0.20 |
| **Junk Food Freq** | 0.15 |

**5. Discussion**

**Key Findings**

1. **BMI is the strongest predictor** of health risk (65% importance).
2. The model performs best in classifying **"low" risk** (F1-score: **0.83**).
3. **"Medium" risk cases are harder to classify** (F1-score: **0.69**).

**Limitations**

* **Imbalanced dataset** (more "medium" cases than "low").
* **Small dataset size** (only 100 samples).
* **Potential overfitting** due to limited data.

**6. Conclusion**

* The **Random Forest model** achieves **~76% accuracy** in predicting health risk levels.
* **BMI is the most influential factor**, followed by exercise and diet.
* **Improvement Suggestions**:
  + Collect more data to balance classes.
  + Try other models (e.g., **XGBoost, SVM**).
  + Include additional features (e.g., age, blood pressure