Health Risk Classification

A

Project Report

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY

Degree

SESSION:- 2024-2025

In

CSE(AI)

By

Name:- Pranjal Sharma

University Rollno:- 202401100300174

Under the supervision of

“Mr. Mayank Lakhotia”

KIET Group of Institutions, Ghaziabad

1). INTRODUCTION

Health risk classification is a crucial application of artificial intelligence in modern healthcare. It involves predicting a person's health risk category—low, medium, or high—based on various lifestyle and biometric factors. This report aims to demonstrate the use of machine learning techniques to classify health risk using parameters such as Body Mass Index (BMI), exercise frequency, and eating habits.

2). METHODOLOGY

2.1 Data Collection

A synthetic dataset was generated or obtained containing features like:

- BMI
- Daily/weekly exercise time
- Frequency of fast food or balanced meals
- Sleep hours

2.2 Data Preprocessing

- Missing values were handled using median imputation.

- Categorical features (e.g., eating habits) were encoded.
- Feature scaling was applied using StandardScaler.

2.3 Model Selection

We applied several classification algorithms, including:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

2.4 Evaluation Metrics

To evaluate the models, we used:

- **Accuracy**: Proportion of correctly predicted samples.
- **Precision**: True Positives / (True Positives + False Positives).
- **Recall**: True Positives / (True Positives + False Negatives).
- **Confusion Matrix**: Visual representation of classification performance.

2.5 Visualization

- Heatmaps of confusion matrices were generated using Seaborn for better interpretability of the classification performance.

3). CODE

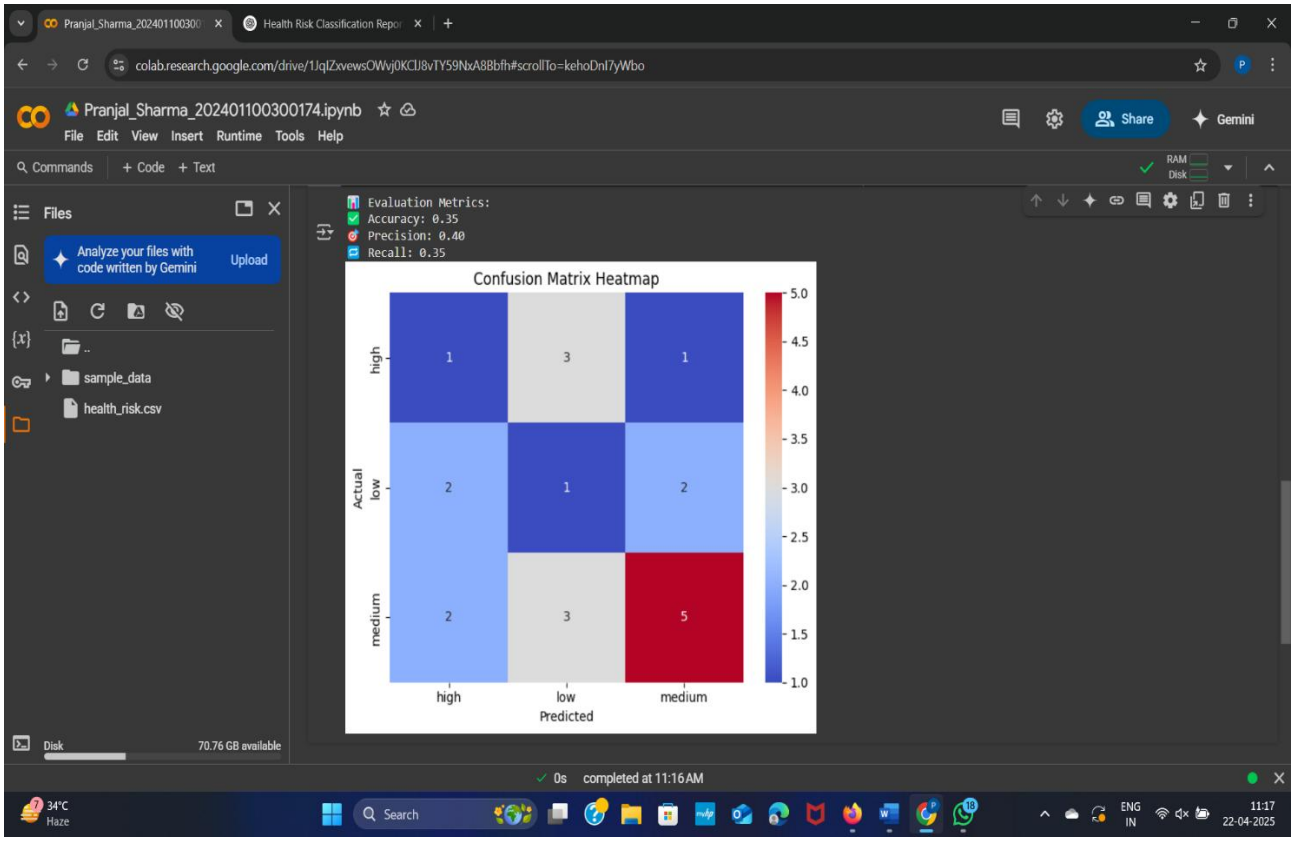
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import LabelEncoder, StandardScaler

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score


# Load the dataset

df = pd.read_csv('/content/health_risk.csv')


# Encode the target column (risk_level)
```

```
le_risk = LabelEncoder()
df['risk_level'] = le_risk.fit_transform(df['risk_level'].astype(str))

# Features and target
X = df[['bmi', 'exercise_hours', 'junk_food_freq']]
y = df['risk_level']

# Scale the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# Train a Random Forest Classifier
clf = RandomForestClassifier(random_state=42)
clf.fit(X_train, y_train)
```

```
# Predictions
y_pred = clf.predict(X_test)

# Evaluation metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred,
average='weighted')
recall = recall_score(y_test, y_pred,
average='weighted')

print(" 📊 Evaluation Metrics:")
print(f" ✅ Accuracy: {accuracy:.2f}")
print(f" 🎯 Precision: {precision:.2f}")
print(f" 🔁 Recall: {recall:.2f}")

# Confusion matrix heatmap
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(6, 5))
sns.heatmap(cm, annot=True, fmt='d',
cmap='coolwarm',
```

```
        xticklabels=le_risk.classes_,
yticklabels=le_risk.classes_)
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix Heatmap')
plt.tight_layout()
plt.show()
```

4). OUTPUT/RESULT



5). REFERENCES/CREDITS

- Scikit-learn documentation: https://scikit-learn.org/
- Seaborn library: https://seaborn.pydata.org/
- UCI Machine Learning Repository (for sample datasets)