Title Page

Project Title: Health Risk Classification
Student Name: Aditya Mishra
Roll Number: 202401100300017
Course: Bachelor of Technology
Institution: KIET Group of Institutions

Introduction

In the modern healthcare landscape, predicting individual health risks using data-driven methods can help in early diagnosis and prevention of diseases. This project aims to classify individuals into health risk categories (low, medium, or high) based on features such as Body Mass Index (BMI), exercise habits, and junk food consumption. Machine Learning models are particularly suitable for such classification problems, where multiple factors contribute to the final decision. This project uses a supervised learning approach with Random Forest Classifier.

Methodology

1. **Data Collection:** The dataset (health_risk.csv) includes BMI, exercise hours per week, junk food frequency, and the target variable risk_level.
2. **Data Preprocessing:** Label encoding was applied to convert the target variable into numeric values. Features were selected and separated as inputs and output.
3. **Model Training:** Data was split into training and testing sets using train_test_split (70%-30%). A Random Forest Classifier was trained.
4. **Evaluation:** Model performance was evaluated using accuracy, precision, recall, and F1-score. A confusion matrix heatmap was also generated.

Code

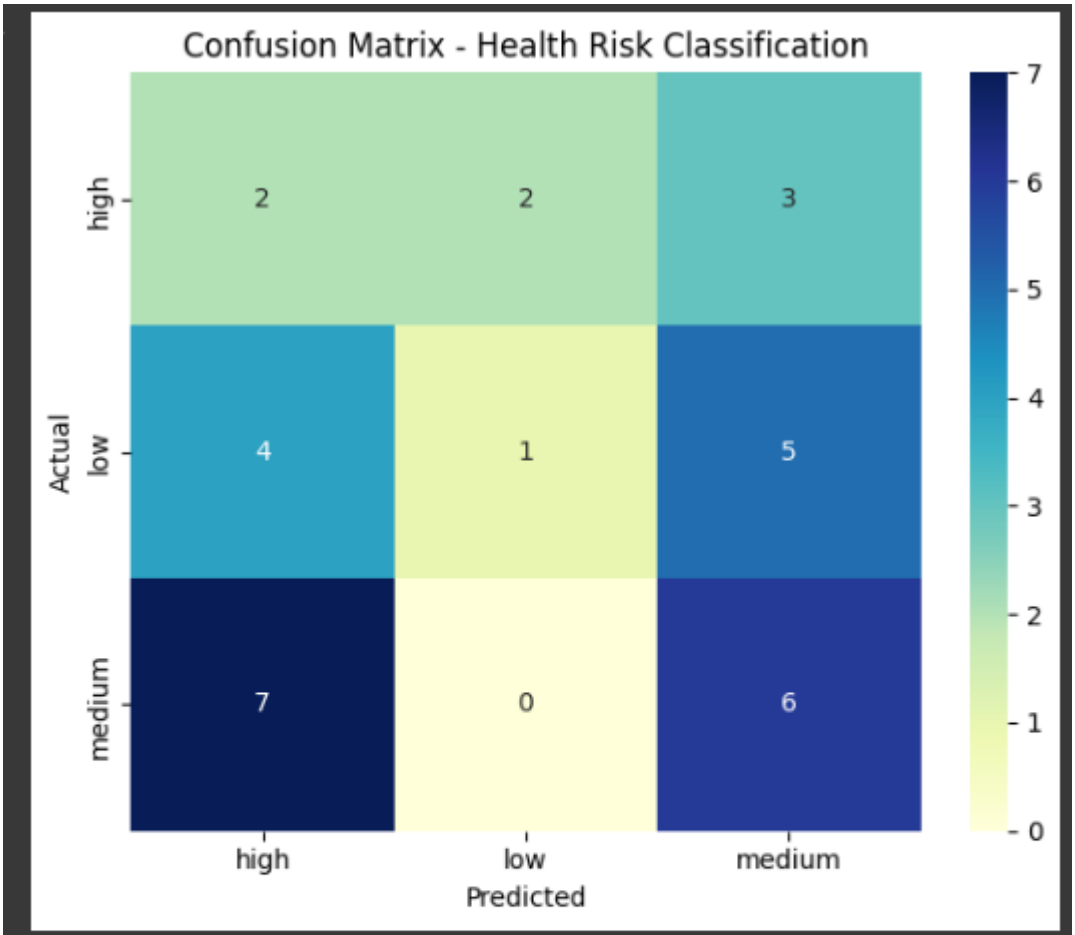
```
import pandas as pd
from sklearn.model_selection import
train_test_split
from sklearn.ensemble import
RandomForestClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import confusion_matrix,
classification_report
import seaborn as sns
import matplotlib.pyplot as plt

file_name = "/content/health_risk.csv"
df = pd.read_csv(file_name)

df.head()

le = LabelEncoder()
df['risk_encoded'] =
le.fit_transform(df['risk_level'])
```

```
X = df[['bmi', 'exercise_hours', 'junk_food_freq']]
y = df['risk_encoded']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

clf = RandomForestClassifier(random_state=42)
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)

cm = confusion_matrix(y_test, y_pred)
labels = le.classes_

plt.figure(figsize=(6, 5))
sns.heatmap(cm, annot=True, fmt='d', cmap='YlGnBu', xticklabels=labels, yticklabels=labels)
plt.title("Confusion Matrix - Health Risk Classification")
plt.xlabel("Predicted")
```

```
plt.ylabel("Actual")
plt.tight_layout()
plt.show()

print("Classification Report:\n")
print(classification_report(y_test, y_pred, target_names=labels))
```

Output/Result

- A confusion matrix heatmap showing classification performance.
- Classification report with:
  - Accuracy
  - Precision
  - Recall
  - F1 Score for each class (low, medium, high)



Classification Report:

	precision	recall	f1-score	support
high	0.15	0.29	0.20	7
low	0.33	0.10	0.15	10
medium	0.43	0.46	0.44	13
accuracy			0.30	30
macro avg	0.31	0.28	0.27	30
weighted avg	0.33	0.30	0.29	30

References/Credits

- **Dataset:** Provided for academic use (health_risk.csv)
- **Libraries:** scikit-learn, pandas, seaborn, matplotlib
- Google Collab platform for implementation and testing
- OpenAI ChatGPT for code explanation and structuring assistance