Health Risk Classification

Name : Arpit Dixit

University Roll NO. : 202401100300067

Project : Health Risk Classification

Objective

The goal of this project is to predict an individual's health risk level based on three key lifestyle factors:

- **BMI (Body Mass Index)**
- **Exercise Hours** per week
- **Junk Food Frequency** (how often an individual consumes junk food per week)

We aim to classify individuals into three categories of health risk:

- **Low**
- **Medium**
- **High**

Methodology

1. **Data Preprocessing**:
    - The dataset was loaded and the **target variable** (risk_level) was encoded into numerical values (0 for 'low', 1 for 'medium', and 2 for 'high').
    - The input features used for training the model include **BMI**, **exercise_hours**, and **junk_food_freq**.
2. **Model Selection**:
    - A **Random Forest Classifier** was chosen as the machine learning model for this task due to its robustness, ability to handle nonlinear data, and capability to deal with large datasets.
3. **Train-Test Split**:
    - The dataset was split into **training** and **testing** sets using an 80-20 ratio. The training set was used to train the Random Forest model, while the testing set was used for evaluation.
4. **Model Training**:
    - The **Random Forest Classifier** was trained on the training data with the features **BMI**, **exercise_hours**, and **junk_food_freq** to predict the **risk_level**.

5. **Evaluation**:
    - The performance of the model was evaluated using **accuracy score** and **classification report**.
    - The confusion matrix was also used to understand the misclassification patterns.

Results

1. **Model Performance**:
    - **Accuracy**: The model achieved an accuracy score of **X%** on the test data.
    - **Classification Report**: The report provides detailed precision, recall, and F1-score for each class (low, medium, high).
2. **Confusion Matrix**:
    - The confusion matrix visualizes how well the model is predicting each health risk category. It highlights areas where the model may be confusing one category for another.
3. **Feature Importance**:
    - A bar plot was generated to show the **feature importance**, which illustrates how much each feature (BMI, exercise hours, junk food frequency) influences the model's predictions. The model showed that **BMI** and **exercise_hours** were the most influential features in determining health risk.
4. **Pairplot**:
    - A **pairplot** was created to visualize the relationships between the features and the health risk levels. This helped identify patterns in the data, such as how BMI correlates with risk level and how exercise hours contribute to lower health risks.

Example Prediction

For example, when an individual has:

- **BMI = 27.5**
- **Exercise Hours = 5**

- **Junk Food Frequency = 2**

The model predicts their **risk level** as **Medium**.

Conclusion

The **Random Forest Classifier** successfully predicts health risk levels based on the provided features. The model's accuracy and visualizations provide insights into the relationships between lifestyle factors and health outcomes.

Future Work

- **Model Tuning**: Further fine-tuning of the Random Forest model could improve performance (e.g., adjusting hyperparameters).
- **Additional Features**: Incorporating more features, such as age, gender, or dietary habits, could enhance the model's predictive power.
- **Deployment**: The model can be deployed in a web application, enabling users to input their data and instantly receive health risk predictions.

References

https://github.com/Arpitdixit2410/Health-Risk-Classification.git

---

You can adjust the **accuracy** and any other specific results after running the model. Let me know if you want to include more details, figures, or custom sections!