Predicting Obesity Levels Based on Eating Habits and Physical Condition

Overview

This Jupyter Notebook aims to predict obesity levels by analyzing various factors such as eating habits, physical condition, and personal attributes. It leverages two classification models: K-Nearest Neighbors (KNN) and Logistic Regression. The goal is to identify which model performs better in classifying instances of obesity and to interpret the significant features contributing to these predictions.

Dataset

The dataset used is **"Estimation of Obesity Levels Based on Eating Habits and Physical Condition"**, sourced from the UCI Machine Learning Repository. It includes features related to:

- Age, gender, and personal habits
- Eating frequency (e.g., vegetable and snack consumption)
- Physical activity frequency
- Other personal metrics, such as family history of overweight or smoking habits

The dataset can be found here.

Models and Techniques

1. K-Nearest Neighbors (KNN)

- KNN is used as a baseline model to classify individuals into obesity categories.
- The optimal value of K is determined through cross-validation, with K = 4 yielding the highest accuracy of 92%.

2. Logistic Regression

- Recursive Feature Elimination (RFE) is applied to select the most important features contributing to obesity prediction.
- The logistic regression model is evaluated based on accuracy and F1 score, and the identified features provide insights into obesity risk factors.

3. Model Evaluation

- Both models are evaluated using accuracy, precision, recall, and F1 score.
- Confusion matrices are generated to assess the classification performance for different obesity levels.

Notebook Structure

1. Data Preprocessing

The dataset is loaded and cleaned, with necessary transformations and encodings applied to prepare it for model training.

2. KNN and Logistic Regression Training

The notebook trains both KNN and Logistic Regression models on the dataset, using cross-validation to determine the best-performing hyperparameters.

3. Feature Importance Analysis

The RFE method is applied to the Logistic Regression model to identify the most important features related to obesity. These include age, snack frequency, physical activity, and family history of overweight.

4. Results and Conclusion

- **KNN** achieved the highest accuracy (92%) and F1 score (0.89), making it the most effective model for this task.
- **Logistic Regression** also performed well but was slightly outperformed by KNN.
- Key features contributing to the predictions are discussed in detail, providing insights into obesity risk factors.

Conclusions

- **KNN Model**: Achieved better accuracy and F1 score than Logistic Regression, making it the preferred model for obesity prediction in this dataset.
- **Logistic Regression**: Identified important features contributing to obesity, including frequency of snacks, vegetable consumption, family history of overweight, and physical activity frequency.
- **Optimal K for KNN**: K = 4 was found to be the optimal number of neighbors, yielding the best model performance.
-

Files Included

- `48622613_Portfolio4.ipynb` - The Jupyter Notebook with all the code and analysis.
- `data/` - A folder containing the dataset ObesityDataSet_raw_and_data_sinthetic.csv

Usage

To explore the analysis, open the notebook in Jupyter Notebook or a compatible environment and execute the cells in sequence. The results will include model performance metrics, confusion matrices, and an analysis of important features.

Dependencies

- Python 3.x
- Pandas
- Scikit-learn
- Jupyter Notebook

To install the required dependencies, run:

```
pip install pandas scikit-learn jupyter
```

In [ ]: