<https://www.kaggle.com/datasets/jayitabhattacharyya/estimation-of-obesity-levels-uci-dataset>

Using the "Estimation of Obesity Levels" dataset, here are some potential topics you could study based on the tasks you mentioned:

1. \*\*Classification\*\*:

- \*\*Obesity Level Prediction\*\*: Use classification models to predict obesity levels based on features like age, eating habits, and physical activity. You could explore binary classification (obese vs. non-obese) or multi-class classification (e.g., different levels of obesity).

- \*\*Health Risk Classification\*\*: Classify individuals into health risk categories (e.g., low, medium, high risk) based on their obesity level and lifestyle factors.

2. \*\*Clustering\*\*:

- \*\*Lifestyle-based Clustering\*\*: Use clustering algorithms to group individuals with similar lifestyle factors (e.g., eating habits, physical activity, alcohol consumption) to uncover hidden patterns.

- \*\*Behavioral Segmentation\*\*: Cluster individuals based on behavioral traits such as physical activity, diet, and habits to identify common health profiles.

- \*\*Obesity Progression Clustering\*\*: Group individuals by common patterns in the progression or development of obesity.

3. \*\*Link Prediction\*\*:

- \*\*Predicting Co-occurrence of Health Conditions\*\*: Using link prediction methods to identify the likelihood of co-occurring health conditions (e.g., obesity linked with diabetes or heart conditions).

- \*\*Linking Lifestyle Factors to Obesity Levels\*\*: Predict the likelihood that certain lifestyle factors (e.g., high fast food consumption, low physical activity) lead to obesity. You could model this as predicting the "links" between lifestyle choices and obesity levels.

Each of these tasks can lead to deeper insights into the relationship between lifestyle habits and obesity, which is critical for health-based interventions and policy recommendations.

Setting a performance target value for your data mining project can be a challenging but crucial task. To back it up with references from competitions or academic papers, here’s how you can approach it:

### 1. \*\*Use Benchmarks from Similar Datasets or Competitions:\*\*

Look for benchmarks or performance scores from competitions that use similar datasets or address similar problems (obesity prediction, health classification, etc.). Kaggle competitions or other machine learning challenges can be good sources for benchmarks.

- \*\*Example Sources\*\*:

- Kaggle Competitions: Search for any health-related datasets that deal with obesity or classification of medical data and look at the winning models’ performance (e.g., accuracy, AUC, F1-score).

- \*\*Example Performance Metrics\*\*: In health-related datasets, you might typically see accuracy scores around 70-85% for classification tasks or clustering evaluation metrics (e.g., silhouette score, Davies-Bouldin index).

### 2. \*\*Look at Academic Papers\*\*:

Search for academic papers that have worked on obesity-related classification or clustering using similar datasets. These papers usually report performance results (e.g., accuracy, precision, recall, F1-score, AUC, clustering scores) that you can use as a reference.

- \*\*Search Resources\*\*:

- \*\*Google Scholar\*\*: Search for keywords like "obesity level prediction classification" or "clustering lifestyle and health patterns."

- \*\*Research Papers\*\*: The UCI dataset itself might have been used in previous research papers, so checking references on UCI or Kaggle for papers linked to this dataset can provide good targets.

- \*\*Example Performance from Papers\*\*:

- A typical classification model predicting obesity might achieve around 70-90% accuracy, depending on the features and complexity of the model.

- In clustering, achieving a high silhouette score (closer to 1) would be desirable but depends on how distinct the clusters are in your data.

### 3. \*\*Realistic Targets Based on Baseline Models\*\*:

Another approach is to start with baseline models (e.g., logistic regression for classification or k-means for clustering) and evaluate their performance. You can then set your target as improving upon these baselines.

- \*\*Example Performance\*\*: A baseline logistic regression model on a health classification task might give an accuracy around 60-70%. Your goal could be to improve this by 5-10% using advanced techniques.

### 4. \*\*References from Competitions or Papers\*\*:

- \*\*Kaggle Health Competitions\*\*: For example, in Kaggle competitions like "Personalized Medicine: Redefining Cancer Treatment" or other healthcare-related datasets, top models can reach very high accuracies, often 80-90%.

- \*\*Papers\*\*: Papers in medical or health informatics journals (e.g., "Journal of Medical Internet Research") often report precision, recall, and F1-scores in the range of 75-90% for well-optimized models. You can use these as a benchmark for your project.

### Suggested Target Values:

- \*\*Classification\*\*: Aim for at least 75-85% accuracy or an F1-score in a similar range, based on what’s achievable in health-related data.

- \*\*Clustering\*\*: A silhouette score above 0.6 can be a good target, though it can vary based on the dataset.

- \*\*Link Prediction\*\*: Look for AUC-ROC scores or precision-recall curves, aiming for an AUC of 0.8 or higher if the relationships between variables are strong.

Would you like help finding specific papers or competition results for references?