Project Outline

Data Mining Group 11

Can Synthetic Data Augmentation Unlock Superior Classification Performance on Small Datasets, or Does It Introduce More Noise Than Signal?

Petar Malamov Iven Beck Danail Ignatovski
Emre Iyigün Thu-Huong Vu Denis Maxheier

March 30, 2025

1 Problem Description

In recent decades, global economic growth has led to significant lifestyle changes, contributing to a rising prevalence of obesity. As countries develop and incomes rise, dietary habits shift towards higher consumption of processed foods, sugary beverages, and fast food. Additionally, urbanization and technological advancements have led to increasingly sedentary lifestyles, reducing physical activity levels. These changes have contributed to an alarming increase in obesity rates worldwide, making it a pressing public health issue. Since the dataset exhibits class imbalance and sparsity in certain obesity categories, we incorporate synthetic data to enhance representation and improve model performance.

This project aims to explore how the inclusion of synthetic data, can improve the performance of various classification algorithms. We will investigate which algorithms benefit most from synthetic data augmentation, especially in the context of imbalanced multiclass classification tasks like obesity risk prediction.

2 Data

Two datasets will be used in this project.

Original dataset consists of **2111** records of real data. The 80/20 split will be used for this dataset with 80% train set and 20% test set.

Synthetic dataset (provided by the Kaggle competition) consists of **34597** records and is generated using deep learning model trained on the original dataset.

Both datasets have same **17 independent features**, with the target variable **NObeyesdad**. All the features are listed in a table on Imgur. The target variable is a multi-class categorical label representing 6 different obesity risk levels/categories. Each of the categories have been defined by a BMI Range.

3 Problem Solving

3.1 Required Preprocessing Steps

Handling missing data is not required for the chosen dataset, since there are not missing values and no duplicate rows.

Feature encoding will be performed in the following way, keeping numerical values in the encoding they are in and encoding categorical variables according to Table 1.

Encoding	Feature
Numerical	Age, Height, Weight, FCVC, NCP,CH20,FAF,TUE,NObeyesdad, CAEC, CALC
One-hot	MTRANS
Binary	Gender, FAVC, Smoke, SCC, family_history_with_overweight

Table 1: Feature encodings

Numerical feature normalization will be applied to FCVC, NCP, CH20, FAF, and TUE to account for differences in their respective value ranges. This step will ensure that variables with larger magnitudes do not disproportionately influence the model's decision-making process. Additionally, standardization will be performed on Age, Height, and Weight to bring these features to a common scale. Given the potential for Height and Weight to introduce multicollinearity and act as false predictors in BMI prediction, a feature selection process will be considered to mitigate their potential impact on model performance.

Handling Class Imbalance During our initial screening, we identified an imbalance in certain categorical variables, including FAVC, CAEC, SMOKE, SCC, and MTRANS. Addressing this imbalance will be crucial to ensuring fair and unbiased model performance. Depending on the specific use case, we will explore various techniques discussed in the lecture, such as resampling methods or algorithmic adjustments, to mitigate the effects of class imbalance.

3.2 Algorithms to be applied to the data

To build an effective multi-class classification model, the following algorithms will be applied:

- **Lazy Classifier as Baseline:** Initially, the LazyClassifier library will be used to benchmark various models and identify promising algorithms. Based on the results from LazyClassifier, selected models will be further fine-tuned. The most likely models to be used are listed below.
- **Decision Tree:** A simple yet interpretable model that serves as a baseline for performance.
- **Random Forest:** An ensemble learning method that improves accuracy and reduces the risk of overfitting by combining multiple decision trees.
- **Support Vector Machine (SVM):** Effective for high-dimensional data and non-linear classification tasks, particularly when the dataset has clear margins of separation.
- **XGBoost and AdaBoost:** Boosting algorithms that sequentially combine weak learners to enhance predictive performance, particularly in handling class imbalances.
- **Artificial Neural Network (ANN):** Explored if the complexity of the problem requires deeper learning techniques to capture non-linear relationships in the data.
- **K-mean Clustering:** Possibly for pattern recognition or feature engineering.
- **Ensemble:** Different Ensemble methods like bagging, boosting and stacking will be tried out.

3.3 Hyperparameter optimization

We begin with **Random Search** for initial tuning, as it is faster than Grid Search and well-suited for models such as Random Forest, Decision Tree, and AdaBoost. Its simplicity makes it ideal for early experimentation.

For more complex models like XGBoost and ANN, we apply **Bayesian Optimization** (via Optuna). This method learns from prior evaluations to efficiently explore the hyperparameter space, reducing the number of required trials while improving performance.

We may eventually use **Grid Search** for models with small hyperparameter spaces, such as SVM, where an exhaustive search remains practical. Across all optimization techniques, we consistently use **Stratified K-Fold Cross-Validation** to ensure class balance and robust performance estimation.

4 Result evaluation

The following methods will be used to evaluate the results:

- **Precision, Recall, and F1-Score** provide a balanced evaluation even in cases of uneven class distribution.
- **Confusion Matrix** will help visualize the performance of the model by showing true positives, true negatives, false positives, and false negatives.
- **Accuracy** may be used as a general measure of model performance; however, its limitations on imbalanced data must be considered.
- **Cost-Sensitive Evaluation**, since this is a health-related problem, the cost of missing a true positive prediction for obesity is higher than incorrect prediction in a non-obese case.
- **R-Squared** (R^2) will be used to evaluate the approach of treating the problem as a regression task.

5 Expected Results

We expect that augmenting real data with synthetic samples will generally improve the performance of classification algorithms when evaluated on original test data. This improvement is attributed to increased input diversity and better coverage of decision boundaries. The augmented data is hypothesized to reduce overfitting and enhance generalization.

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Projektarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und in der untenstehenden Tabelle angegebenen Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT	Rephrasing, More Concise Formulation	section 1, section 2, subsection 3.3, section 5	+

Signature Mannheim, March 30, 2025