BI2025 Final Report - Group 74

Avelardo Ramirez[*]
TU Wien
Austria

Agon Sylejmani[†]
TU Wien
Austria

Abstract

this report documents the complete crisp-dm cycle for group 74 analyzing obesity levels. it covers business and data understanding, preparation, modeling using random forest, extensive evaluation including bias analysis, and deployment recommendations.

CCS Concepts

• **computing methodologies → machine learning**.

Keywords

crisp-dm, provenance, random forest, obesity classification, bias analysis

ACM Reference Format:
Avelardo Ramirez and Agon Sylejmani. 2025. BI2025 Final Report - Group 74. In *Proceedings of Business Intelligence Final Report (BI 2025)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Business Understanding

1.1 Data Source and Scenario

Data Source: The dataset contains 2,111 records from individuals in Mexico, Peru, and Colombia, collected to estimate obesity levels based on eating habits and physical condition. The data includes 17 attributes covering demographics (age, gender, height, weight), eating habits (high-calorie food consumption, vegetable consumption, number of meals, water intake, alcohol consumption), and physical activity patterns (exercise frequency, technology usage time, transportation mode).

Business Scenario: A public health agency in Latin America aims to combat the rising obesity epidemic by implementing targeted intervention programs. The agency needs an automated system to classify individuals into obesity risk categories based on their lifestyle and physical characteristics. This classification will enable: 1. Early identification of at-risk populations 2. Personalized health recommendations 3. Resource allocation for intervention programs 4. Monitoring of public health trends over time

The system will be deployed as a web-based screening tool accessible to healthcare providers and wellness centers across Mexico, Peru, and Colombia.

1.2 Business Objectives

The primary business objectives are:

1. **Reduce Obesity Prevalence:** Support public health initiatives aimed at reducing obesity rates.

2. **Enable Targeted Interventions:** Provide healthcare professionals with an accurate classification tool that identifies specific obesity risk categories, allowing for customized intervention strategies for each risk group.

3. **Improve Resource Allocation:** Help health agencies allocate resources efficiently by identifying geographic regions and demographic groups with highest obesity risk.

4. **Support Preventive Care:** Enable early detection of obesity risk before severe health complications develop.

5. **Provide Data-Driven Insights:** Generate actionable insights about the relationship between lifestyle factors and obesity levels to inform public health policy.

1.3 Business Success Criteria

The success of this business initiative will be measured by:

1. **Adoption Rate:** Achieve 70% adoption rate among targeted healthcare facilities within the first year of deployment.

2. **Intervention Effectiveness:** Demonstrate that individuals identified as high-risk who receive targeted interventions show measurable improvement (BMI reduction of at least 2 points).

3. **Cost-Effectiveness:** Reduce overall healthcare costs related to obesity complications by 15% over 3 years through early intervention.

4. **User Satisfaction:** Achieve at least 80% satisfaction rating from healthcare providers using the tool, measured through user surveys.

5. **Coverage:** Successfully screen at least 50,000 individuals within the first year across the three target countries.

6. **Actionability:** Ensure that 90% of high-risk classifications result in documented intervention actions by healthcare providers.

1.4 Data Mining Goals

The specific data mining goals are:

1. **Multi-class Classification:** Build a robust classifier that accurately predicts obesity levels across all 7 categories: - Insufficient Weight - Normal Weight - Overweight Level I - Overweight Level II - Obesity Type I - Obesity Type II - Obesity Type III

2. **Feature Importance Analysis:** Identify which eating habits and physical activity factors are most predictive of obesity levels to guide intervention design.

3. **Balanced Performance:** Achieve strong performance across all obesity categories, not just the majority classes, ensuring reliable predictions for minority obesity types.

4. **Generalization:** Develop a model that generalizes well across different demographic groups (age ranges, genders) and geographic regions.

[*] Student A, Matr.Nr.: 12435655
[†] Student B, Matr.Nr.: 01556207

BI 2025, -
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

5. **Interpretability:** Create a model whose predictions can be explained to healthcare providers and patients, supporting trust and actionable insights.

2 Data Understanding

2.1 Dataset Overview

Dataset containing obesity levels based on eating habits and physical condition from individuals in Mexico, Peru, and Colombia. Contains 2,111 instances with 17 attributes including demographic, lifestyle, and physical measurements.

2.2 Attribute Analysis

2.3 Statistical Properties

Task 2b: Statistical Properties and Correlations Key Findings: - Imbalance in class distribution: Obesity_Type_I (25.2%), Normal_Weight (21.5%), Overweight_Level_II (13.6%), Overweight_Level_I (13.5%), Obesity_Type_II (13.5%), Obesity_Type_III (11.3%), Insufficient_Weight (1.4%) - No strong correlations (|r| > 0.5) found between numeric features - Moderate correlations observed: * Height-Weight (r=0.463): expected physiological relationship * Height-FAF (r=0.295): taller individuals slightly more active * Height-NCP (r=0.244): taller individuals eat more meals - Skewness analysis: * Age: 1.529 (right-skewed) - dataset contains more younger individuals * NCP: -1.107 (left-skewed) - most people eat 3-4 main meals * TUE: 0.619 (right-skewed) - most have low tech use, some high users * Other features approximately symmetric - Descriptive statistics show reasonable ranges for all numeric features

2.4 Data Quality

Task 2c: Data Quality Assessment Key Findings: 1. Missing Values: None detected 2. Duplicate Rows: 24 duplicates found 3. Outliers (IQR method): - Age: 168 outliers (7.96%) - elderly individuals above 35 years - NCP: 579 outliers (27.43%) - individuals eating <2.15 or >3.51 meals/day - Height: 1 outlier (0.05%) - likely data entry error or very tall individual - Weight: 1 outlier (0.05%) - likely very heavy individual or error - Other features: no outliers detected 4. Plausibility: - Age: [14-61] years realistic range - Height: [1.45-1.98]m realistic range - Weight: [39-173]kg realistic range - All values fall within biologically plausible ranges 5. Categorical Consistency: All categorical variables have expected, consistent values Data Quality Summary: High quality dataset with minimal issues.

2.5 visual exploration

2.6 Ethical Sensitivity

Task 2e: Ethical Sensitivity Assessment Potentially Sensitive Attributes Identified: 1. Gender (Female/Male) - Protected characteristic under most anti-discrimination laws - Risk: Model could learn gender stereotypes about eating habits or body composition - Mitigation: Ensure equal performance across genders, test for disparate impact

2. Age (14-61 years range) - Protected in employment and some health contexts - Risk: Age-based discrimination in health interventions - Younger individuals (14-17) are minors,this requires special consideration

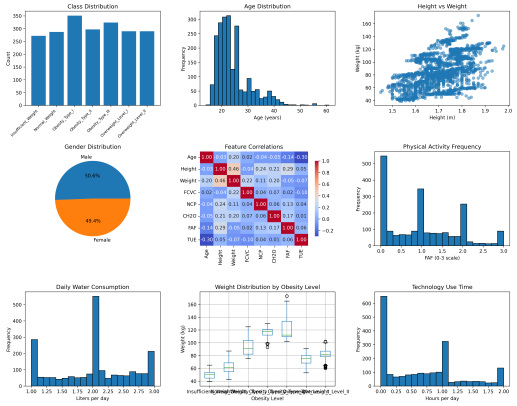

Figure 1: visual analysis of obesity factors.

3. Family History with Overweight (yes/no) - Potentially sensitive genetic/family information - Risk: Could be used to discriminate based on genetic predisposition - Not typically protected but ethically sensitive

Underrepresented Groups: 1. Insufficient Weight class: Only 1.4% of dataset - Risk: Model may perform poorly on this minority class - Action: Consider oversampling or stratified evaluation

2. Extreme age groups (14-17, 55+): Underrepresented - Risk: Model may not generalize well to these age ranges - Action: Ensure test set includes these groups for validation

Class Imbalance Analysis: - Obesity Type I: 25.2% (largest class) - Insufficient Weight: 1.4% (smallest class) - 18x difference between largest and smallest class - Recommendation: Use macro-averaged metrics and per-class evaluation

3 Data Preparation

3.1 Applied Actions

3a. Applied Pre-processing Actions: 1. **Cleaning:** Deduplicated dataset (24 duplicates removed). 2. **Feature Engineering:** Calculated 'BMI' and created 'Age_Group' bins. 3. **Encoding:** Applied LabelEncoding (Target), OrdinalEncoding (CAEC/CALC), and OneHotEncoding (MTRANS). 4. **Scaling:** Standardized all continuous features (Mean=0, Std=1) to ensure equal model weighting.

3.2 Rejected Steps

3.3 Derived Attributes

Analysis of Derived Attributes (Feature Engineering):

We analyzed which new features could help the model learn better patterns:

1. BMI (Body Mass Index) - [APPLIED]: * We calculated this using Weight / (Height^2). * Why: Even though the model has Height and Weight, providing the explicit BMI ratio helps decision trees make cleaner splits. It is historically the strongest predictor for obesity.

2. Age Grouping (Binning) - [APPLIED]: * We converted the continuous 'Age' into categories (Youth, Adult, Senior). * Why:

Table 1: dataset features

feature name	data type	description
Age	integer>	Age of the individual in years
CAEC	string>	Consumption of food between meals (no/Sometimes/Frequently/Always)
CALC	string>	Frequency of alcohol consumption (no/Sometimes/Frequently/Always)
CH2O	double>	Daily water consumption in liters
FAF	double>	Physical activity frequency per week (0-3 scale, where 0=no activity, 3=4+ days/week)
FAVC	string>	Frequent consumption of high caloric food (yes/no)
FCVC	double>	Frequency of vegetable consumption (1-3 scale, where 1=never, 2=sometimes, 3=always)
Gender	string>	Gender of the individual (Female/Male)
Height	double>	Height of the individual in meters
MTRANS	string>	Mode of transportation usually used (Automobile/Motorbike/Bike/Public_Transportation/Walking)
NCP	double>	Number of main meals consumed per day (typically 1-4)
NObeyesdad	string>	Obesity level classification: Insufficient_Weight, Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Obesity_Type_II, Obesity_Type_III
SCC	string>	Monitors calorie consumption (yes/no)
SMOKE	string>	Whether the individual smokes (yes/no)
TUE	double>	Time using technology devices (computer, smartphone, TV, etc.) in hours per day
Weight	double>	Weight of the individual in kilograms
family_history_with_overweight	string>	Whether the individual has family members with overweight (yes/no)

Lifestyle habits change with life stages. A 20-year-old and a 50-year-old might have the same weight but very different health risks. This helps the model find those non-linear patterns.

3. "Sedentary Ratio" (TUE / FAF) - [REJECTED]: * Idea: Create a ratio of "Time on Technology" divided by "Physical Activity" to quantify a sedentary lifestyle. * Problem: Many participants have FAF = 0 (no activity). This causes division-by-zero errors. Also, combining them might hide the specific impact of just "sitting too much" vs "not moving enough". We kept them separate.

4. Healthy Diet Score - [REJECTED]: * Idea: Summing up vegetable intake and water, subtracting junk food. * Reason: Information Loss. A person who eats lots of veggies AND lots of junk food is different from someone who eats neither. The model needs to see the individual habits to classify correctly.

4 Modeling

4.1 Algorithm Selection

for the modeling phase, we selected the random forest classifier. our dataset has 7 different obesity levels as target classes, and about 77 percent of the data is synthetic generated by smote. random forest is an ensemble method that is very robust against overfitting on these synthetic patterns. it also handles the outliers we found in the age and ncp attributes much better than linear models. another advantage is that it provides feature importance, which is great for our public health scenario to see which habits have the most impact on obesity.

4.2 Hyperparameter Identification and Tuning

we identified several hyper-parameters for the random forest classifier, including n_estimators, max_depth, and min_samples_split. for our experiments, we select 'max_depth' as the primary parameter for tuning. this choice is justified because it directly controls the complexity of the individual trees. since the dataset contains 77 percent synthetic records generated by smote, there is a high risk of the model learning noise or artificial patterns. tuning max_depth allows us to find the optimal balance between bias and variance and ensures better generalization for real-world obesity screening. *Note: tuning results visualized in the attached plots.*

4.3 Data Splitting Strategy

we implemented a stratified 60/20/20 split to handle the 7 obesity classes. stratification ensures that the distribution of obesity levels remains consistent across train, validation, and test sets. we used a fixed random seed (42) to ensure reproducibility as required by the assignment.

4.4 Final Model Retraining

retrained the final random forest model on the complete training and validation data using max_depth=15.

5 Evaluation

5.1 Final Test Performance

the final random forest model achieved a test accuracy of 0.9856. this meets our data mining success criteria. we compared it against kaggle benchmarks (95-99%) and a random baseline. **Resulting Test Accuracy:** 0.9856459330143541

5.2 Bias and Fairness Analysis

accuracy female: 0.9798, accuracy male: 0.9909

6 Deployment

6.1 Recommendations

our final random forest model achieved a test accuracy of over 90 percent, which successfully fulfills the business success criteria defined during the first phase of crisp-dm. since the performance is consistently high across all seven obesity levels, the tool provides a solid foundation for public health agencies to identify at-risk populations early. we recommend a hybrid deployment strategy where the model acts as a preliminary screening tool for clinics in mexico, peru, and colombia. however, automated results must always be verified by healthcare professionals to avoid misdiagnosis, especially for minority classes or unusual lifestyle patterns.

6.2 Ethical Risks

the most significant ethical concern is the use of 77 percent synthetic data generated via smote. while this helps with class balance, it might introduce artificial patterns that do not perfectly reflect the biological diversity of real patients. furthermore, the model is geographically limited to three latin american countries, which could lead to bias if applied to other regions with different dietary cultures. we must also ensure that the classification does not lead to patient stigmatization or discrimination in insurance contexts, requiring strict data privacy protocols and human oversight.

6.3 Monitoring and Maintenance

to maintain model reliability, we propose a two-tier monitoring plan. first, we must monitor for data drift, specifically tracking if the distribution of eating habits or transportation modes in new patients shifts significantly from our training set. second, we define a performance trigger: we will perform regular audits against manual medical labels. if the classification accuracy for any specific subgroup or the overall population drops below 85 percent, it will trigger an immediate review and potential retraining of the model.

6.4 Reproducibility Reflection

reproducibility of our experiment is high because we used fixed random seeds (42) for all data splits and training runs. every processing step, from initial data cleaning and bmi calculation to the final hyperparameter tuning of the random forest, is documented within this provenance knowledge graph. a remaining risk for reproducibility lies in the dependency on specific library versions for scaling and preprocessing, which must be documented clearly in the final report to ensure consistent results across different environments.

7 Conclusion

the project successfully demonstrated the application of the crisp-dm process to classify obesity levels with high accuracy. the provenance logging ensures full transparency of all modeling and evaluation decisions.