ICIT Journal (Innovative Creative and Information Technology) | Print ISSN: 2356-5195 |
Online ISSN: 2654-8704 | Vol. 12 No. 2, August 2026 | DOI Prefix:
https://doi.org/10.33050/icit.v12i2.4431

The Application of Business Intelligence for Analysis and Prediction of Obesity Rates using The Decision Tree Method

Elsa Mutiarasari*[1], Dina Auliya Rahmadani[2], Sandro Alfeno[3]
[1,2,3]University of Raharja, Tangerang, Indonesia
*Corresponding author: elsa@raharja.info

Article Information:
Received: 15 June 2026
Revised: 16 July 206
Accepted: 22 August 2026
DOI: 10.33050/icit.v12i2.4431

ABSTRACT
Obesity is a health issue influenced by diet, physical activity intensity, and lifestyle. This study aims to apply Business Intelligence (BI) to analyze and estimate the prevalence of obesity using the RapidMiner-based Decision Tree method and interactive visualization through Google Looker Studio. The dataset used consists of the Obesity Dataset, containing 1,610 entries and using 15 attributes. The model was tested using a 10-fold cross-validation approach and successfully achieved an accuracy of 71.55% ± 3.18%. The results of this study emphasize that height, age, and physical activity are the most important factors in obesity classification. In addition, the Google Looker Studio dashboard can present interactive data visualization, allowing users to easily explore obesity patterns related to lifestyle and health conditions. This study suggests that the combination of Decision Tree and Business Intelligence can increase efficiency in health data analysis.

Keywords—Business Intelligence, Decision Trees, Obesity, RapidMiner, Google Looker Studio.

1. INTRODUCTION

Obesity is a health condition characterized by excessive body fat accumulation that can increase the risk of various chronic diseases such as type 2 diabetes, hypertension, cardiovascular problems, stroke, and certain types of cancer. According to data from the World Health Organization, the prevalence of obesity continues to surge worldwide, making it a major public health issue today. The causes of obesity are highly complex and involve poor dietary habits, lack of physical activity, a sedentary lifestyle, genetic factors, and a tendency to spend excessive time using technology.

Advances in information technology and data science are opening up new opportunities to explore health issues through the application of machine learning and Business Intelligence (BI). BI enables the collection, processing, analysis, and visualization of data in an interactive manner. As a result, the findings can serve as a basis for making more informed decisions. In the context of health, BI can play a role in identifying behavioral patterns among the public that are associated with the risk of obesity.

One of the machine learning methods frequently used for data classification is the Decision Tree. This algorithm has the advantage of generating classification rules that are simple, easy to understand, and easy to visualize. In addition, Decision Trees can handle both numerical and categorical data without requiring complex normalization processes. Therefore, this method is well-suited for use in obesity analysis, which encompasses various aspects of an individual's lifestyle and physical condition.

Several previous studies have analyzed obesity mapping using methods such as K-Nearest Neighbor, Support Vector Machine, Random Forest, and Decision Tree. However, most of these studies have focused solely on comparing the effectiveness of these algorithms without integrating the analysis results into an interactive Business Intelligence platform. Furthermore, the entire process—from machine learning to dashboard visualization—is still rarely carried out.

Regarding this issue, this study aims to apply Business Intelligence in analyzing and estimating

obesity levels using the Decision Tree method based on RapidMiner and visualizing it through Google Looker Studio. It is hoped that this research can produce an accurate classification model and provide an interactive dashboard that allows individuals without a technical background to better understand obesity patterns in a clearer and more informative way.

2. RESEARCH METHOD

2.1 Dataset

This study relies on a dataset released by Suleyman Alpaslan Sulak in 2024 titled "Obesity Dataset." The dataset used in this study was obtained from the Kaggle platform and is publicly available. The dataset includes medical records and lifestyle data related to obesity, comprising a total of 1,610 rows and 15 columns of attributes, with no missing values across all attributes.

This dataset consists of 14 predictor attributes and 1 target attribute. The predictor attributes include: Gender, Age, Height, and Family History of Obesity.

Body Weight, Fast Food Consumption, Frequency of Vegetable Consumption, Number of Main Meals per Day, Snacking Between Meals, Smoking, Daily Fluid Intake, Calorie Counting, Physical Activity, Duration of Technology Use, and Type of Transportation Used. All of these attributes have been encoded in numerical format as integers.

The target attribute is a categorical class divided into four categories: Class 1 (Normal Weight) with 73 records, Class 2 (Class I Overweight) with 658 records, Class 3 (Class II Overweight) with 592 records, and Class 4 (Obesity) with 287 records. This class distribution indicates an imbalance that is important to consider when performing modeling.

Table 1. Data Distribution by Obesity Class.

Class	Label	Number of Records
1	Normal Weight	73
2	Overweight Level 1	658
3	Overweight Level 2	592
4	Obese	287
Total		**1.610**

2.2 Processing Flow in RapidMiner

The Decision Tree modeling process is carried out using the RapidMiner Studio platform, a visual data science and AI software tool. This platform allows users to prepare data, perform machine learning, deep learning, text mining, and predictive analysis without having to write code (no-code/low-code). Workflows are created visually through a drag-and-drop interface that links various workflows developed in the research.

ICIT Journal (Innovative Creative and Information Technology) | Print ISSN: 2356-5195 |
Online ISSN: 2654-8704 | Vol. 12 No. 2, August 2026 | DOI Prefix:
https://doi.org/10.33050/icit.v12i2.4431

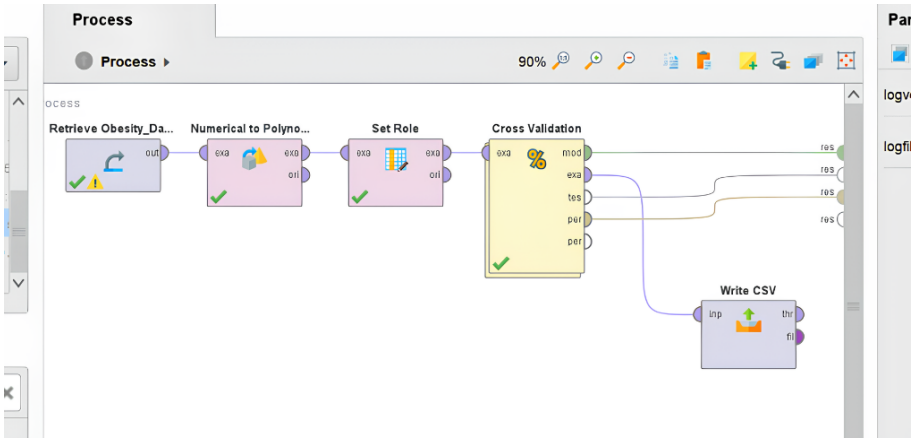

Figure 1. Decision Tree Modeling Workflow in RapidMiner

The workflow consists of four main operators connected in sequence. Here's an explanation of each operator:

1. Retrieve Obesity Dataset: This operator is used to retrieve the Obesity dataset from the local RapidMiner repository into the process. The dataset consists of 1,610 rows and 15 attribute columns.
2. Numerical to Polynomial: This operator converts numeric attributes (specifically, target or class attributes) into polynomial-type attributes (nominal or categorical). This conversion is necessary because the Decision Tree algorithm in RapidMiner requires label attributes to be in nominal format in order to perform multiclass classification.
3. Set Role: This operator is used to assign roles to each attribute in the dataset. The Class attribute is set as the label (target), while the other 14 attributes are set as regular (predictor features).
4. Cross Validation: This operator performs a cross-validation process (k-fold cross-validation), which includes the sub-processes of training a Decision Tree model and evaluating its performance. The output of this process consists of a trained model, test data (Example Set), and performance metrics, which are then passed to the process output.

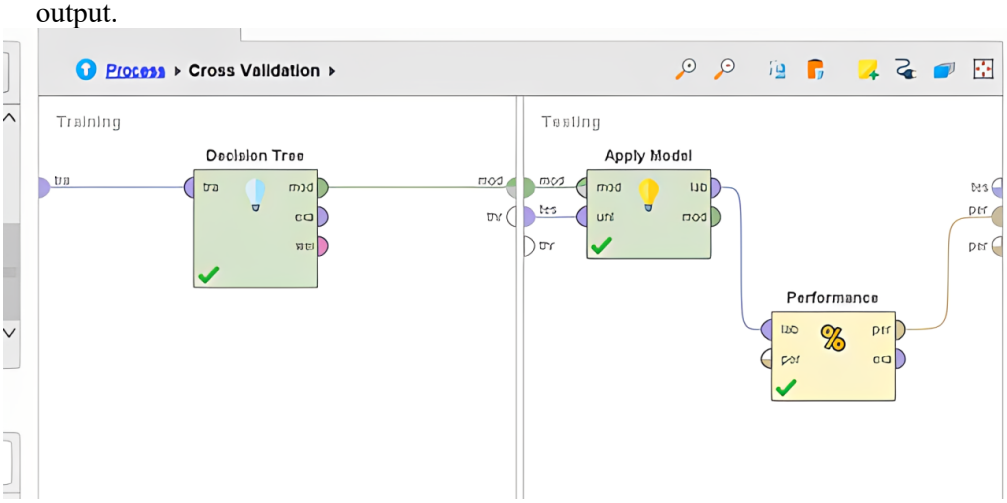

Figure 2. The Cross-Validation Sub-Process in RapidMiner

On the Training side, the Decision Tree operator receives training data (tra) and generates a trained model (mod), which is then passed to the Testing side. On the Testing side, the Apply Model operator uses the trained model on the test data (tes) to generate prediction labels (lab). Next, the Performance operator calculates the classification evaluation metrics by comparing the prediction results and the actual labels. The performance output (per) is passed out of the Cross Validation process to be presented as the final result.

ICIT Journal (Innovative Creative and Information Technology) | Print ISSN: 2356-5195 | Online ISSN: 2654-8704 | Vol. 12 No. 2, August 2026 | DOI Prefix: https://doi.org/10.33050/icit.v12i2.4431

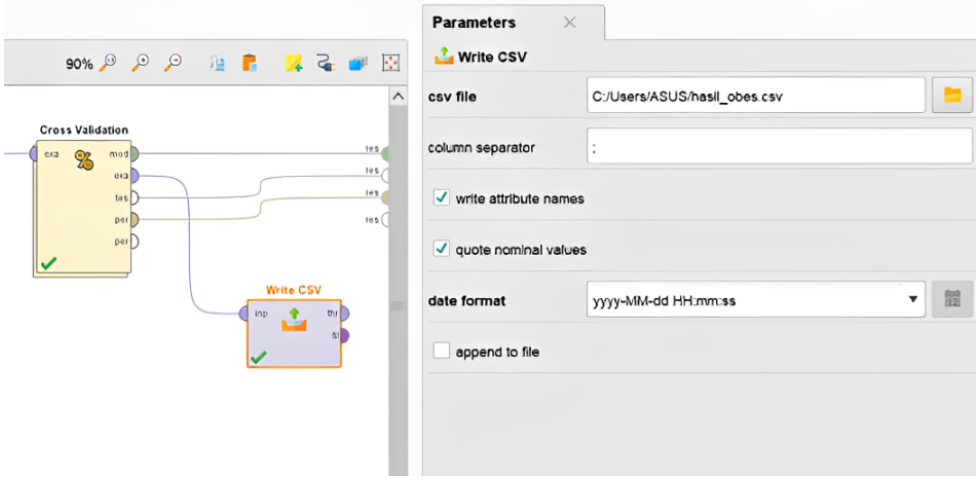

Figure 3. The Process of Saving Model Evaluation Results Using the Write CSV Operator in RapidMiner.

In the "Write CSV" configuration, the column separator used is a semicolon (;) in accordance with regional settings, with the options to include attribute names and enclose numerical values in quotes enabled. The date format used is yyyy-MM-dd HH:mm:ss.

2.3 Parameter Decision Tree

The method used in this study is the Decision Tree, with attribute selection criteria based on Information Gain. The Decision Tree operates by repeatedly splitting the dataset based on the attribute that yields the highest reduction in uncertainty (entropy) at each node.

The parameters set for the Decision Tree operator in RapidMiner are as follows: (1) Criterion: Information Gain — used as a quality measure for selecting split attributes; (2) Maximal Depth: 10 — sets the maximum depth limit of the decision tree to prevent overfitting; (3) Confidence: 0.25 — the confidence-based pruning threshold for removing insignificant branches; (4) Minimal Gain: 0.01 — the minimum gain value a split must achieve to be considered relevant; (5) Minimum Leaf Size: 2 — the minimum number of records that must be present in a leaf node; (6) Minimum Size for Split: 4 — the minimum number of records in a node to allow for further splitting; and (7) Number of Prepruning Alternatives: 3 — the number of alternatives considered in the pre-pruning process.

The Decision Tree process generates decision trees with classification rules that are immediately understandable to business users, making it an ideal method for Business Intelligence applications focused on predictive analytics.

2.4 Model Validation Using Cross-Validation

The model was validated using the k-fold cross-validation method, which is integrated into the Cross Validation operator in RapidMiner. This method randomly divides the dataset into k non-overlapping subsets (folds). In each iteration, one fold is used as the test set, while the remaining k-1 folds are used as the training set. This process is repeated k times until each fold has been used as the test set once.

The value of k used in this study is k = 10 (10-fold cross-validation). Choosing k = 10 is a common practice in machine learning that is considered to provide stable performance estimates with low variance. The metrics used for evaluation include Accuracy (overall accuracy), Precision, Recall, and F1-Score for each obesity category. The average of the results from the 10

ICIT Journal (Innovative Creative and Information Technology) | Print ISSN: 2356-5195 | Online ISSN: 2654-8704 | Vol. 12 No. 2, August 2026 | DOI Prefix: https://doi.org/10.33050/icit.v12i2.4431

iterations is used as an estimate of the model's performance on new data (generalization performance).

Cross-validation was chosen to avoid bias that might arise from a single split of the training and test data, and to maximize the use of all available data during the training and evaluation process.

2.5 Visualization with Google Looker Studio

The prediction results file (hasil_obes.csv) has been uploaded to Google Looker Studio as a data source for creating an interactive Business Intelligence dashboard. The dashboard is designed with two pages: the first page presents an overview of the predictions and the distribution of obesity classes along with their risk factors, while the second page focuses on the analysis of respondents' lifestyles. Interactive filters that group data by gender, age, obesity category, and physical activity level are provided so that non-technical users can explore the data on their own.

3. RESULTS AND DISCUSSION

3.1 Decision Tree

The decision tree model obtained from the training process using 10-fold cross-validation produces a relatively complex tree structure. Due to the large size of the tree, the visualization is divided into six sections (Figures 4–9).

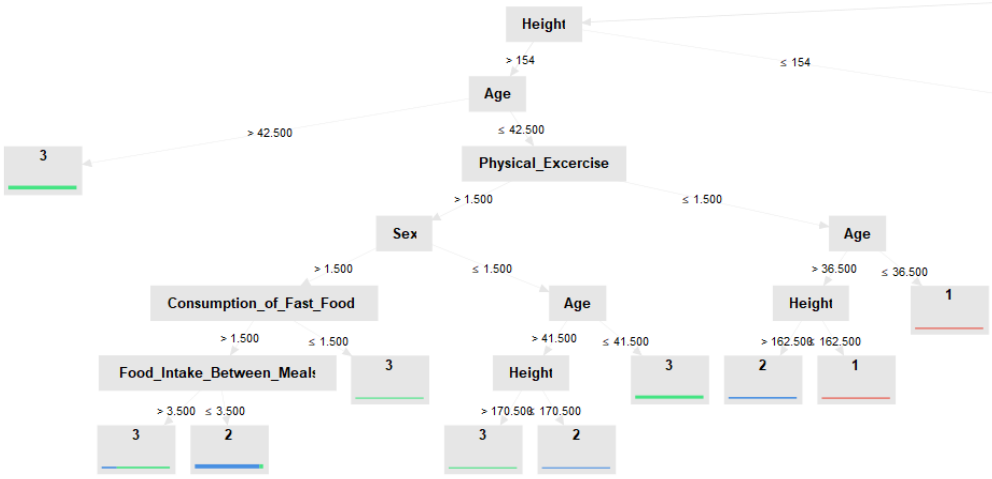

Figure 4. Visualization of the Decision Tree for Obesity Prediction.

Figure 4 shows the top (root) of the decision tree. The root node starts with the Height attribute, with thresholds of ≤ 154 cm and > 154 cm, indicating that height is the attribute with the highest information gain. In the Height > 154 branch, the Age attribute serves as the next separator with a threshold of 42.5 years. Individuals > 42.5 years are immediately grouped into class 3 (Overweight Level II), indicating a strong relationship between advanced age and a high level of overweight. Younger individuals are further evaluated based on Physical_Exercise.

ICIT Journal (Innovative Creative and Information Technology) | Print ISSN: 2356-5195 |
Online ISSN: 2654-8704 | Vol. 12 No. 2, August 2026 | DOI Prefix:
https://doi.org/10.33050/icit.v12i2.4431

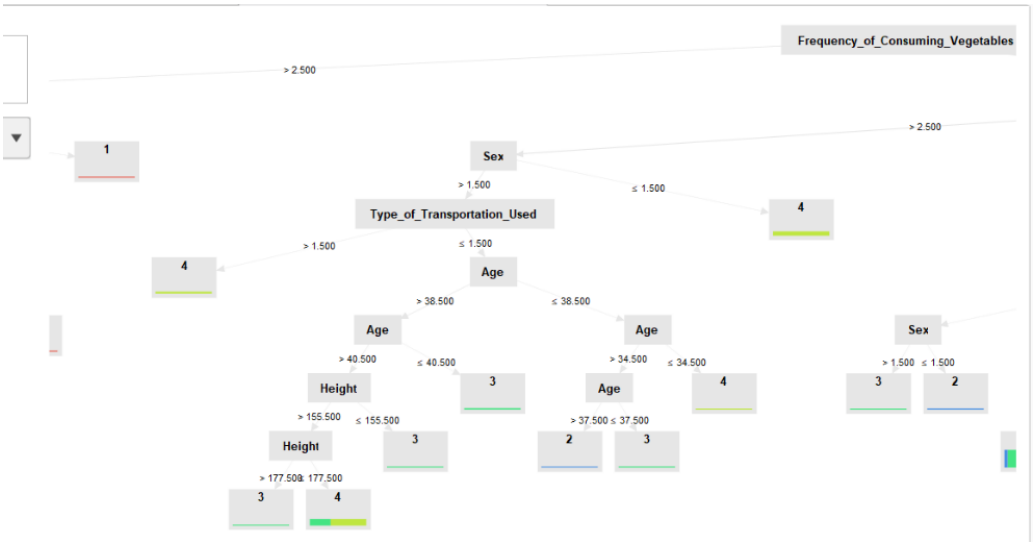

Figure 5. The Deepest Level of a Decision Tree.

Figure 5 shows the deepest level of the decision tree, which includes Frequency_of_Consuming_Vegetables, Number_of_Main_Meals_Daily, Physical_Exercise, Smoking, Type_of_Transportation_Used, Sex, Age, and Height. The combination of low frequency of vegetable consumption, the number of meals consumed, and smoking habits results in a leaf classified as Level 4 (Obese). Overall, Height, Age, and Physical_Exercise are the three most important attributes for classification at the top level, while other lifestyle attributes serve to distinguish at deeper levels.

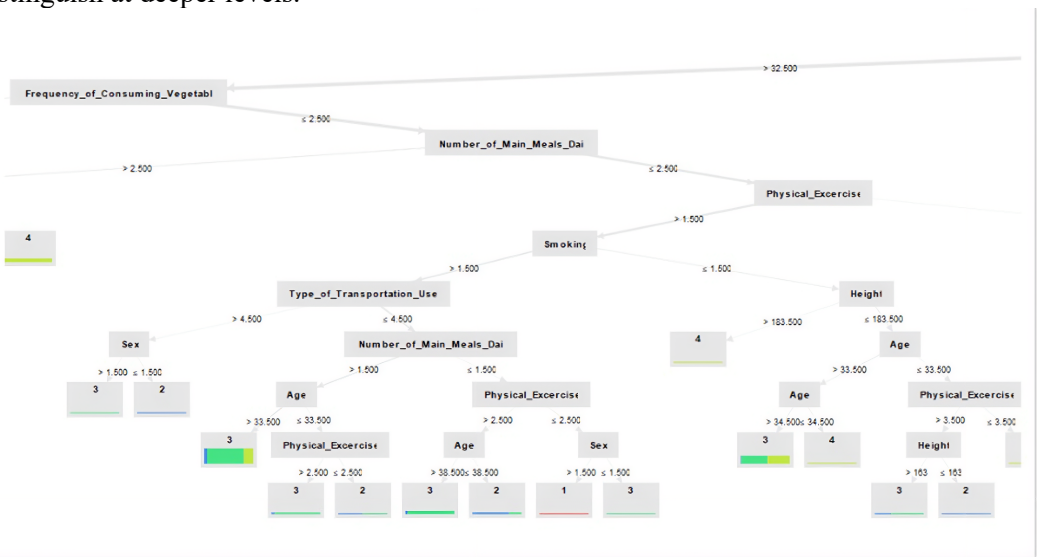

Figure 6. Cluster Analysis by Age and Sex.

Figure 6 shows a branch centered on Age, with sub-branches including Schedule_Dedicated_to_Technology, Frequency_of_Consuming_Vegetables, Smoking, Physical_Exercise, and Sex. The emergence of Schedule_Dedicated_to_Technology as a boundary indicates that a sedentary lifestyle (high duration of technology use) is associated with an increased risk of obesity.

ICIT Journal (Innovative Creative and Information Technology) | Print ISSN: 2356-5195 | Online ISSN: 2654-8704 | Vol. 12 No. 2, August 2026 | DOI Prefix: https://doi.org/10.33050/icit.v12i2.4431

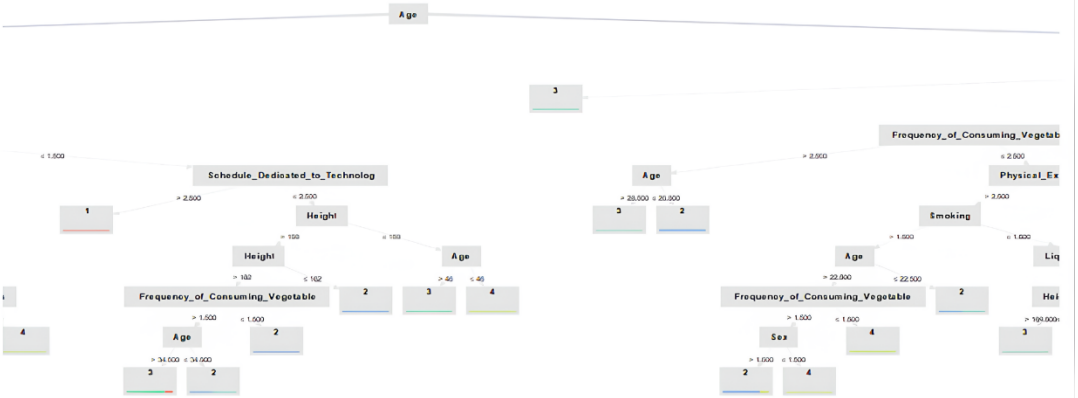

Figure 7. Lifestyle- and Family History-Focused Categories in the Classification of Obesity

Figure 7 shows a branch involving Frequency_of_Consuming_Vegetables, Physical_Exercise, Number_of_Main_Meals_Daily, Smoking, Liquid_Intake_Daily, and Overweight_Obese_Family. The presence of a family history of obesity indicates that genetic factors also play a role in classification under certain conditions, while Number_of_Main_Meals_Daily reflects the influence of daily eating habits on obesity levels.

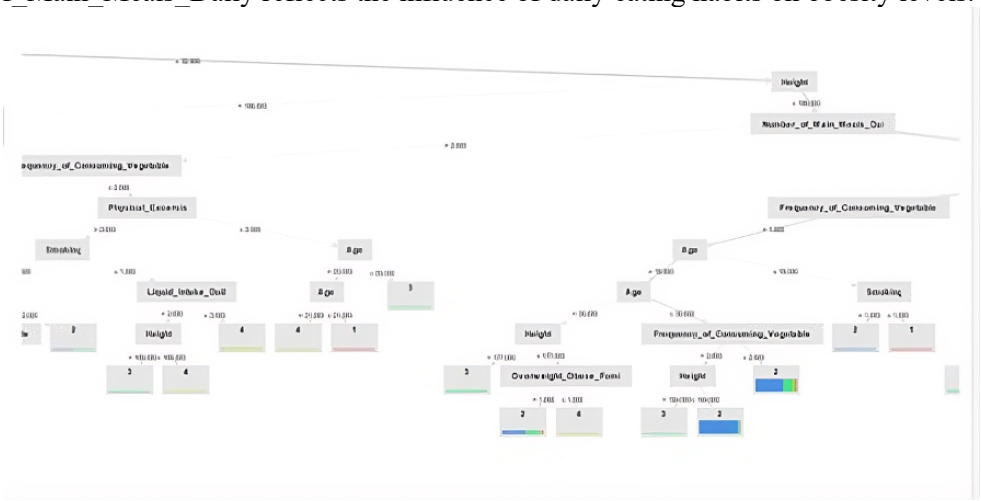

Figure 8. Categories Focused on Lifestyle and Age in the Classification of Obesity.

Figure 8 shows a branch involving Physical Exercise, Frequency of Consuming Vegetables, Age, Type of Transportation Used, and Smoking. The presence of the Smoking and Type of Transportation Used attributes at this level indicates that overall lifestyle—rather than a single factor—plays a role in determining obesity class.

ICIT Journal (Innovative Creative and Information Technology) | Print ISSN: 2356-5195 |
Online ISSN: 2654-8704 | Vol. 12 No. 2, August 2026 | DOI Prefix:
https://doi.org/10.33050/icit.v12i2.4431

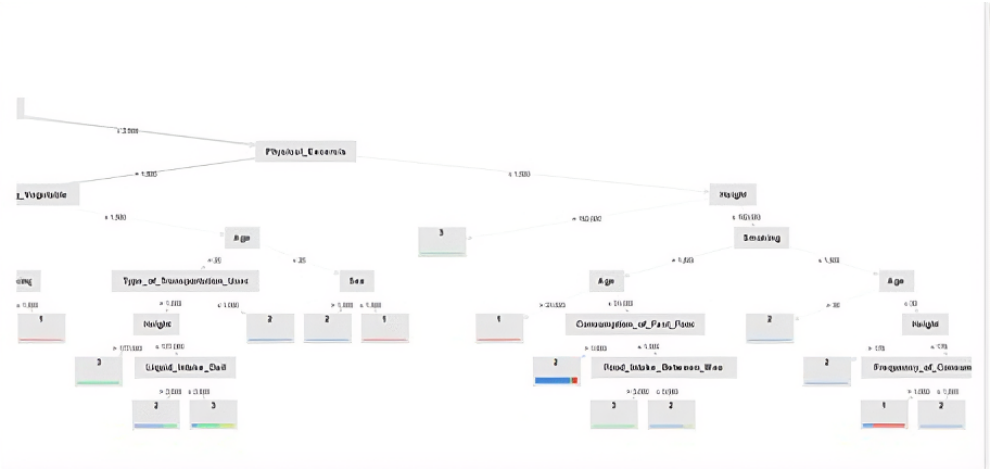

Figure 9. The Physical_Exercise branch.

In the Physical_Exercise category ≤ 1.5 (low physical activity), the next classification is based on age, with a threshold of 36.5 years. Individuals who rarely exercise, are > 36.5 years old, and are > 162.5 cm tall tend to be classified into class 2 (Overweight Level I), indicating that low physical activity combined with age and height plays a major role in increasing the risk of obesity.

3.2 Confusion Matrix and Model Performance Evaluation

accuracy: 71.55% +/- 3.18% (micro average: 71.55%)

	true 2	true 3	true 4	true 1	class precision
pred. 2	562	124	18	40	75.54%
pred. 3	76	402	108	3	68.25%
pred. 4	6	60	160	2	70.18%
pred. 1	14	6	1	28	57.14%
class recall	85.41%	67.91%	55.75%	38.36%	

Figure 10. Evaluation of the Decision Tree Model's Performance.

The Decision Tree model achieved a total accuracy of 71.55% ± 3.18% (micro-average: 71.55%). The standard deviation of 3.18% indicates that the model's performance was consistent across all 10 folds. Class 2 (Overweight Level I) recorded the highest-class recall (85.41%) with 562 accurate predictions out of 658 original data points and a precision of 75.54%—this model is most effective at identifying this class due to its dominant sample size. Class 3 (Overweight Level II) showed a recall of 67.91% and a precision of 68.25%. Class 4 (Obese) showed a recall of 55.75% and a precision of 70.18%. Class 1 (Normal Weight) showed the lowest performance, with a recall of 38.36% and a precision of 57.14%.

This means that only about 38 out of every 100 individuals who are of normal weight were correctly identified by the model, while the remainder were mostly misclassified into the neighboring Overweight Level I category. This weakness is directly attributable to the severe class imbalance in the dataset, where Class 1 accounts for only 73 of the 1,610 records (4.5%), compared to 658 records (40.9%) for Class 2. With so few training examples, the Decision Tree struggles to form split rules that specifically isolate Normal Weight cases and instead tends to favor the decision boundaries of the majority classes.

In practical terms, this indicates that the model is more reliable for detecting overweight and obese cases than for confirming a normal weight status. This is a limitation that should be considered before the model is used to support clinical or public health decisions. Addressing this limitation—for example, through oversampling techniques such as SMOTE, class-weighted

ICIT Journal (Innovative Creative and Information Technology) | Print ISSN: 2356-5195 |
Online ISSN: 2654-8704 | Vol. 12 No. 2, August 2026 | DOI Prefix:
https://doi.org/10.33050/icit.v12i2.4431

learning, or collecting additional data for the minority class—is therefore an important direction for improving the model in future work.

Table 2. Results of the Classification Model Evaluation.

Class	Precision	Recall	Description
1 – Normal Weight	57,14%	38,36%	Lowest (Minority)
2 – Overweight Level I	75,54%	85,41%	Best
3 – Overweight Level II	68,25%	67,91%	Pretty Good
4 - Obese	70,18%	55,75%	**Average**
Overall Accuracy (Micro Average)	**71,55% ± 3,18%**		

3.3 Business Intelligence Visualizations in Google Looker Studio

The results from the Decision Tree model, which were exported in CSV format, were then visualized using Google Looker Studio. The dashboard was designed with two pages to present comprehensive analytical information in a way that is easy to understand for users without a technical background.

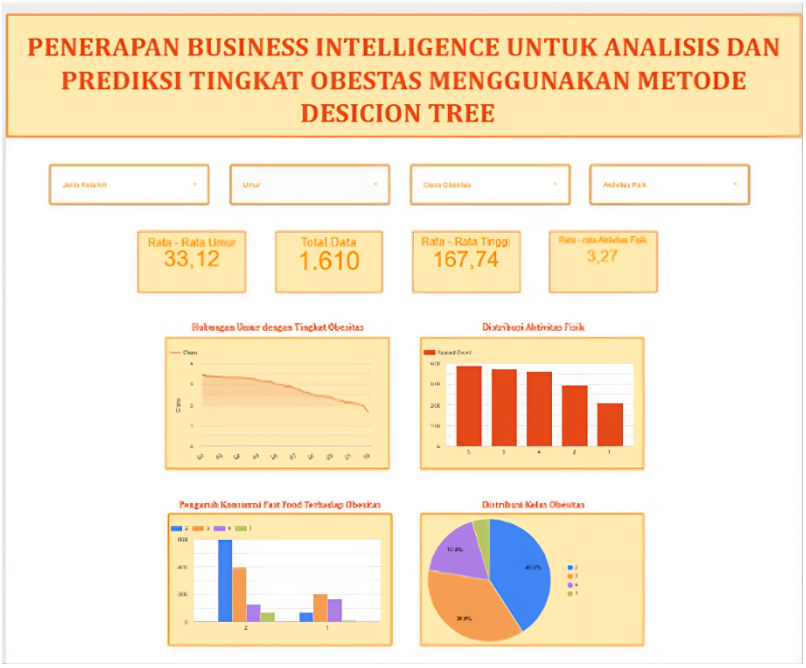

Figure 11. Main Dashboard View Using Google Looker Studio

The dashboard display contains a summary of descriptive statistics and a visual representation of the class distribution of the prediction results. The scorecard shows that the average age of respondents is 33.12 years, with an average height of 167.74 cm and an average physical activity score of 3.27. The total number of data points visualized is 1,610 records. The line chart showing the relationship between age and obesity level indicates that Category 3 (Overweight Level II) and Category 2 (Overweight Level I) dominate in every age group, while Category 4 (Obese) tends to increase in the 30–45 age range.

The bar chart of physical activity distribution shows that respondents with a physical activity score of 5 (very active) constitute the largest group, yet they are still spread across various

ICIT Journal (Innovative Creative and Information Technology) | Print ISSN: 2356-5195 | Online ISSN: 2654-8704 | Vol. 12 No. 2, August 2026 | DOI Prefix: https://doi.org/10.33050/icit.v12i2.4431

obesity categories, indicating that physical activity is not the sole factor influencing obesity levels. The pie chart of obesity category distribution confirms the prevalence of Class 2 and Class 3 in the data used. An interactive filter that considers gender, age, obesity category, and physical activity level makes it easier for stakeholders to conduct independent analyses as needed.

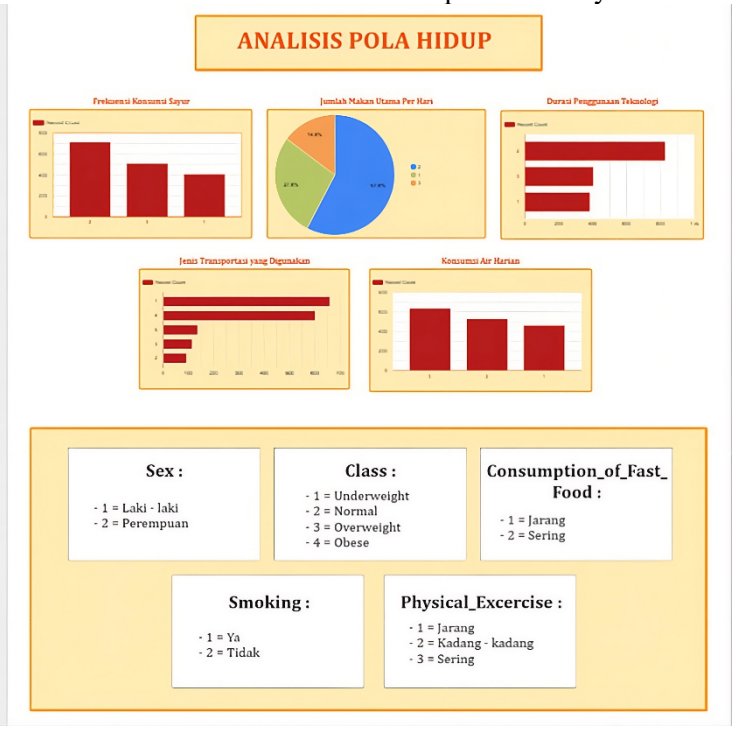

Figure 12. The second page of the RapidMiner dashboard.

The second page displays a dashboard focused on analyzing respondents' lifestyle patterns. The bar chart showing vegetable consumption frequency indicates that the group with Category 2 (moderate) vegetable consumption has the highest proportion. The pie chart showing the number of daily main meals indicates that many respondents (57.6%) eat twice a day, followed by 27.6% who eat once a day, and only 14.8% who eat three times a day. This pattern is significant because irregular eating patterns can lead to fat accumulation. The horizontal bar chart on technology usage duration shows a prevalence in category 2 (moderate), indicating that many respondents have a relatively sedentary lifestyle. The bar chart on mode of transportation shows that category 1 (likely motorized vehicles) is the most frequently used, contributing to low levels of daily physical activity.

The bar chart of daily water consumption shows a balanced distribution across the three categories, with Category 3 (high) accounting for the largest share. The information panel at the bottom of the dashboard provides a legend of attribute codes to help non-technical users understand the visualization.

Overall, the results of this study indicate that the application of Business Intelligence using Decision Trees with RapidMiner can produce a predictive model for obesity levels with an accuracy of 71.55%. The results from the Looker Studio dashboard support the model analysis by highlighting behavioral and lifestyle patterns that are strongly associated with obesity levels. This BI approach has proven successful in transforming raw data into information that can support decision-making processes in the public health sector.

4. CONCLUSION

Based on the findings of the study, the application of the Decision Tree method using RapidMiner proved effective in classifying obesity levels based on various factors such as lifestyle, physical activity, and individual conditions. The model developed using the 10-fold cross-validation technique demonstrated an accuracy of 71.55% ± 3.18%, indicating that its performance is quite good and consistent in predicting obesity across different classes.

Analysis of the results shows that the attributes of height, age, and physical activity are the most influential factors in the obesity classification process. In addition, other attributes such as vegetable consumption frequency, smoking habits, technology use, and family history of obesity also play a role in the formation of classification rules in the decision tree.

The integration of Business Intelligence via Google Looker Studio successfully presents analysis results in the form of interactive dashboards that are easy to understand for users without a technical background. These visualizations are highly effective in simplifying the data interpretation process and supporting data-driven decision-making in the public health sector. Thus, this study demonstrates that the combination of machine learning and Business Intelligence can be effectively utilized to analyze and visualize obesity patterns in an informative and interactive manner.

For future research, it is recommended to apply different classification methods, such as Random Forest, XGBoost, or ensemble learning, to improve model performance. Additionally, class imbalance can be addressed to optimize classification performance for the minority class.

5. SUGGESTION

Based on the findings from our previous research, future studies could employ alternative methods such as Random Forest or XGBoost to improve the model's accuracy. Additionally, increasing the volume of data and incorporating a wider variety of attributes could also be considered to make obesity prediction results more effective. From a Business Intelligence perspective, dashboards could be designed to be more interactive to support decision-making processes in the field of public health.

REFERENCES

[1] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, 2021.

[2] Dwitamara et al., "Analisis Perbandingan Klasifikasi Obesitas Menggunakan Metode K-Nearest Neighbor dan Ensemble Learning," Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research, 2024. [Online].

[3] E. Prasiwiningrum, "Projection and Visualization of Health Workforce Data in Indonesia Using Google Looker Studio (2015–2017)," Journal of ICT Applications and System (JICTAS), vol. 2, no. 2, pp. 74–80, 2023. [Online].

[4] E. Turban, R. Sharda, and D. Delen, Decision Support and Business Intelligence Systems, 10th ed. Pearson Education, 2018. ISBN: 978-0-13-301067-2.

[5] M.Yhogha Ismail Ibn Ibrahim & Sandi Badiwibowo Atim "Klasifikasi Level Obesitas Menggunakan Decision Tree C4.5 dalam Menentukan Akurasi pada Kriteria Information Gain, Gain Ratio, Gini Index," JIKI: Jurnal Ilmu Komputer dan Informatika, vol. 5, no. 2, 2024. [Online].

[6] M. A. R. Saputra, "Penerapan Business Intelligence untuk Menganalisis Data Kasus Covid-19 di Provinsi Jawa Barat Menggunakan Platform Google Data Studio," Jurnal Ilmiah Komputasi, vol. 22, no. 2, pp. 187–196, 2023.

[7] N. Koklu and S. A. Sulak, "Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals' Social and Physical Activities," Sinop University Journal of Natural Sciences, vol. 9, no. 1, pp. 217–239, Jun. 2024.

[8] S. A. Utiarahman, A. Mulawati, and M. M. Pratama, "Analisis Perbandingan KNN, SVM, Decision Tree dan Regresi Logistik untuk Klasifikasi Obesitas Multi Kelas," KLIK: Kajian Ilmiah Informatika dan Komputer, vol. 4, no. 6, pp. 3137–3146, 2024.
[9] S. A. Utiarahman, A. Mulawati, and M. M. Pratama, "Penerapan Support Vector Machine dan Random Forest Classifier untuk Klasifikasi Tingkat Obesitas," Jurnal FASILKOM, vol. 14, no. 3, 2024.
[10] World Health Organization (WHO), "Obesity and Overweight," WHO Fact Sheet, 2024.
[11] Y. J. Sihite, A. V. Saragih, and J. A. Aziz, "Desain dan Implementasi Algoritma Decision Tree untuk Klasifikasi Kategori Obesitas," JRIIN: Jurnal Riset Informatika dan Inovasi, vol. 3, no. 12, pp. 3040–3048, 2026. [Online].
[12] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012. ISBN: 978-0-12-381479-1.
[13] RapidMiner GmbH, "Altair RapidMiner Documentation," 2025.
[14] Suparto Darudiato, S. W. Santoso, and S. Wiguna, "Business Intelligence: Konsep dan Metode," CommIT (Communication and Information Technology) Journal, vol. 4, no. 1, pp. 63–67, 2010.
[15] N. R. Ni'mah, A. Larasati, C. N. Laksana, and E. Mohamad, "Business Intelligence System Design Based on Performance Monitoring Dashboard Using Online Analytical Processing (OLAP) Method," Jurnal Ilmiah Teknik Industri, vol. 23, no. 2, 2024.