APPLICATION OF XGB CLASSIFIER FOR OBESITY RATE PREDICTION

Duwi Cahya Putri Buani1, Nia Nuraeni2*

Informatika
Universitas Nusa Mandiri
duwi.dcp@nusamandiri.ac.id 1, nia.nee@nusamandiri.ac.id 2*
(*) Corresponding author

ABSTRACT

According to the Ministry of Health, the percentage of the population in Indonesia who are overweight is 13.5% for adults aged 18 years and over, while 28.7% are obese with BMI>=25 and obese with BMI>=27 as much as 15.4%. Meanwhile, at the age of children 5-12 years, 18.8% were overweight and 10.8% were obese. From these data, early detection of obesity levels is needed. From these data, prevention is needed so that the percentage of the population who experience obesity can decrease, one of the efforts that can be done is to do early detection of obesity, to do early detection of obesity can be done using Machine Learning. In this study, it was discussed about the prediction of obesity levels using 7 (seven) models, namely Naive Bayes (NB), Random Forest (RF), K-NN, Decision Tree Classifier (DTC), SVM, XGB Classifier (XGB), Logistic Regression (LR) from the seven models used to predict the obesity level of XGB Classifier (XGB) which has the highest accuracy, namely Accuracy 0.96, with an f1-score of 0.96, Precision and recall 0.96.

Keywords: Obesity; XGB Classifier; CRIPS-DM

INTRODUCTION

Overweight or obesity is an abnormal or excessive accumulation of fat that threatens health. According to the WHO standard Body Mass Index (BMI) if it is greater than 25 then it is considered overweight, and if the Body Mass Index (BMI) is greater than 30 then it is considered obese (Fuadi & Meidian, 2017; James et al., 2020; Jiang et al., 2016; Setiyo Nugroho et al., 2018). Obesity is considered an epidemic, with more than 4 million deaths per year from being overweight or obese in 2017, in line with the global burden of disease (WHO, 2023). Overweight and obesity are on the rise in adults and children. From 1975 to 2016, the prevalence of overweight or obese children and adolescents aged 5 to 19 worldwide more than quadrupled, from 4% to 18% (Sherly & Sitanggang, 2022). Obesity is one aspect of the malnutrition problem, with more people obese than underweight in all regions except Africa and Asia. Once considered only a problem in high-income countries, overweight and obesity are now rising dramatically in low- and middle-income countries, especially in urban areas. Most overweight or obese children live in developing countries, where growth rates are 30%
higher than in developed countries. From the description above, it is necessary to have early detection of obesity so that the population in Indonesia and even in the world can prevent endemic obesity or obesity.

The degree of obesity can be determined based on eating habits and physical condition consists of several parameters that can be used to measure the accuracy of predictions of obesity rates (Setiyani Et Al., 2023).

One way to do early detection is to predict obesity levels using Machine Learning Algorithms. Machine Learning is a method that is often used in research, Machine learning is used to improve the results of prediction or automatic detection (Hasanah & Nurheppah, 2023; Rolhan et al., 2019; Supriyadi et al., 2020; Yusa & Sindu, 2017).

In this study the author uses the XGB Classifier (XGB) algorithm; this study is expected to prevent the increase in the number of people who are obese by measuring and improving lifestyle. The limitation in this study is only up to finding the best model for early detection of obesity.

Previous research with Obesity Rate Prediction has been conducted using several algorithms including research using the KNN classification algorithm with an accuracy result of 79.96%, (Sibi & Widiarti, 2005) using a naive bayes algorithm the accuracy results obtained 80%, using the SVM algorithm the accuracy results obtained 71.80%. From previous research, it can be seen that the accuracy results are still not optimal so it is necessary to use other models that can produce more accurate accuracy. (Muslehatin & Ibnu, 2017; Sitanggang & Sherly, 2022)

RESEARCH METHODS

In this study the author used the CRISP-DM method, the CRISP-DM method is still very capable of being used to conduct research (Kurniawan Et Al., 2023). The stages of the CRISP-DM method include Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment (Fadillah, 2015). The Flow Chart of the CRIP-DM Method can be seen in figure 1.

RESULTS AND DISCUSSION

1. Business Understanding

According to the Ministry of Health, the percentage of the population in Indonesia who are overweight is 13.5% for adults aged 18 years and
over, while 28.7% are obese with BMI>=25 and obese with BMI>=27 as much as 15.4%. Meanwhile, at the age of children 5-12 years, 18.8% were overweight and 10.8% were obese. From these data, early detection of obesity levels is needed.

2. Data Understanding

The data used is public data which is an estimate of obesity rates in people from Mexico, Peru and Colombia, with ages between 14 and 61 years and diverse eating habits and physical conditions. Then the information is processed so that 17 attributes are obtained, the data amounts to 2111 records. The following is an explanation of the attributes in Table 1.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Respondent’s Gender</td>
</tr>
<tr>
<td>Age</td>
<td>Age of Respondents</td>
</tr>
<tr>
<td>Height</td>
<td>Respondent’s Height</td>
</tr>
<tr>
<td>Weight</td>
<td>Respondent’s Weight</td>
</tr>
<tr>
<td>family_history_with_overweight</td>
<td>Family History of Being Overweight</td>
</tr>
<tr>
<td>FAVC</td>
<td>Frequency consumption of high caloric food</td>
</tr>
<tr>
<td>FCVC</td>
<td>Frequency of consumption of vegetables</td>
</tr>
<tr>
<td>NCP</td>
<td>Number of main meals</td>
</tr>
<tr>
<td>CAEC</td>
<td>Consumption of food between meals</td>
</tr>
<tr>
<td>SMOKE</td>
<td>Smoke</td>
</tr>
<tr>
<td>CH2O</td>
<td>Consumption of water daily</td>
</tr>
<tr>
<td>SCC</td>
<td>Calories consumption monitoring</td>
</tr>
<tr>
<td>FAF</td>
<td>Physical activity frequency</td>
</tr>
<tr>
<td>TUE</td>
<td>Time using technology devices</td>
</tr>
<tr>
<td>CALC</td>
<td>Consumption of alcohol</td>
</tr>
<tr>
<td>MTRANS</td>
<td>Transportation used</td>
</tr>
<tr>
<td>NObeeyesdad (Obesity Rate)</td>
<td>- with data classification using insufficient weight, normal weight, overweight level I, overweight level II, Type I obesity, Type II obesity, type III obesity.</td>
</tr>
</tbody>
</table>

Source: (Palechor & Manotas, 2019)

Of the 17 attributes, there is 1 label, namely NObeeyesdad and there are 7 classifications of which can be seen in figure 2.

In figure 2 of the NObeeyesdad attribute there are 7 classes including Obesity Type 1 there are 351 data, ObesityType 3 there are 324, ObesityType 3 is 297, Overweight Level I is 290, Overweight Level 2 is 290 data, Normal Weight is 287 data and Insufficient Weight is 272 data.

3. Data Preparation

From the data obtained, analysis is carried out by checking the same data, blank data, and transforming the data and dividing the data into testing and trending data by dividing 70% of testing data and 30% of traning data so that the data used is trending 1477 and data for testing 634.

4. Modeling

The modeling process is carried out by making several models to get the best model, the algorithm used in this study is:

a. Naive Bayes (NB)

This method is believed to be useful in providing decision-making recommendations. (Silvana Et Al., 2020)

b. Random Forest (RF)

Random Forest (RF) is a method that can improve the accuracy of results when generating attributes randomly for each node. (Nusrang, 2022)

c. K-NN

The KNN algorithm is a method of classifying data based on the shortest distance to the data object. (Cholil Et Al., 2021)

d. Decision Tree Classifier (DTC)

C45 or also called Decision Tree Algorithm is an algorithm that can be used to form decision trees, decision trees are one method that is very easy to interpret by humans. (Nasrullah, 2021)

e. SVM

Support Vector Machine (SVM) is a technique that can separate two data sets from two different classes by maximizing the limits of the separator function (hyperplane). (Dasmasela Et al., 2022).
f. XGB Classifier (XGB)

XGBoost or Extreme Gradient Boosting is a decision tree learning-based classification method that applies ensemble enhancement techniques. The classification has several methods, one of which is the xXtreme Gradient Boosting (XGBoost) method. In the use of several examples of research using classification techniques, these calculations can produce excellent precision (Rizky Mubarok Et Al., 2022)(Yulianti Et al., 2022)

g. Logistic Regression (LR)

Logistic regression is a method of statistical analysis used to describe the relationship between a dependent variable with two or more categories and one or more independent variables on a categorical scale.(Featured Et Al., 2017), 7 (Seven) The model used to conduct research is a classification algorithm that has the ability to make predictions, the author wants to test the algorithm or model that has the best accuracy. In this study Testing was carried out using a 10-fold cross-validation technique for NB, RF, K-NN, DT, SVM and XGB classification algorithms resulting in matrix confusion, matrix confusion is a method that can be used to measure the performance of classification algorithms. The following is the comparison result of the 7 algorithms used can be seen in figure 3.

Figure 3. Model Accuracy Results

Figure 3 is a graph of the accuracy results obtained after the algorithm is applied to obesity data the results are NB has an accuracy of 43.85%, RF has an accuracy of 95.43%, K-NN has an accuracy of 98.01%, DTC has an accuracy of 93.53%, SVM has an accuracy of 93.15%, XGBoost has an accuracy of 96.69% and LR has an accuracy of 79.34% of these results, the model that has the highest accuracy is XGBoost.

5. Evaluation

The fifth stage of the CRISP-DM method is Evaluation, evaluation is carried out by looking at the Accuracy, Precision, Recall, and F1 Score values. Here’s how the Accuracy, Precision, Recall, and F1 Score are calculated:

\[
\text{Accuracy} = \frac{TP}{\text{jumlah data}}\tag{1}
\]

\[
\text{Precision} = \frac{TP}{TP + FP}\tag{2}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}\tag{3}
\]

\[
F - 1 \text{ Score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}\tag{4}
\]

Because the data used is multiclass data, there are only True Positive (TP), Positive False (FP) and Negative False (FN), which can be used to calculate Accuracy, Precision, Recall and F1 Score. The following are the evaluation results of the 7 (Seven) Models used:

<table>
<thead>
<tr>
<th>Table 2. Model Evaluation Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algo</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>NB</td>
</tr>
<tr>
<td>RF</td>
</tr>
<tr>
<td>K-NN</td>
</tr>
<tr>
<td>DTC</td>
</tr>
<tr>
<td>SVM</td>
</tr>
<tr>
<td>LR</td>
</tr>
<tr>
<td>XGB</td>
</tr>
</tbody>
</table>

Table 2 shows that the Model Using the XGBoost Algorithm has the highest accuracy value where True Positive (TP) with 606, False Negative (FN) of 28, Positive False of 28 Accuracy 0.96, F1-score of 0.96, Precision and recall 0.96. Thus it can be concluded that the Model using XGBoost has better results.

CONCLUSIONS

This study uses the Method CRISP-DM to determine the best model in predicting obesity levels, the CRISP-DM Model has a very clear flow or step so that the results of the study are valid. In this study, the author used 7 (seven) models to predict the level of obesity from 7 (seven) models obtained as follows: Naive bayes with True Positive (TP) with 278, False Negative (FN) with 356, False Positive with 356 Accuracy 0.44, f1-score of 0.44, Precision and recall 0.44, Random Forest with True Positive (TP) with 600, False Negative (FN) with 34, Positive False 34 Accuracy 0.95, f1-score of 0.95, Precision and recall 0.95, K-NN with True Positive (TP) with 558, False Negative (FN) with 76, False Positive with 76 Accuracy 0.88, f1-score of 0.88, Precision and recall 0.88, Decision Tree Classifier with True Positive (TP) with 596, False Negative (FN) with 38, Positive False 38 Accuracy 0.94, f1-score of 0.94, Precision and recall 0.94, SVM with True Positive
(TP) with 337, False Negative (FN) of 297, False Positive of 297 Accuracy 0.53, f1-score of 0.53, Precision and recall 0.53, Logistic Regression with True Positive (TP) with 503, False Negative (FN) of 131, False Positive of 131 Accuracy 0.79, f1-score of 0.79, Precision and recall 0.79, and XGBoost with the highest accuracy value where True Positive (TP) with 606, False Negative (FN) of 28, Positive False of 28 Accuracy 0.96, f1-score of 0.96, Precision and recall 0.96. Thus the best model of the study is XGBoost or XGB Classifier.

REFERENCES


Rizky Mubarok, M., Hertenio, R., Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lambung Mangkurat Jalan Ahmad Yani Km, I, & Selatan, K. (2022). Hyper-


WHO. (2023). *Obesity*. https://www.who.int/health-topics/obesity#tab=tab_1


