Feature Selection Methods in Obesity Prediction: An Experimental Analysis

Aleksandra Sretenović, Marija Đukić [0000-0002-1136-4278], Ana Pajić Simović [0000-0002-9058-8260], Ognjen Pantelić [0000-0002-8925-4976]

University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia
as20233203@student.fon.bg.ac.rs, marija.djukic@fon.bg.ac.rs, ana.pajic.simovic@fon.bg.ac.rs, ognjen.pantelic@fon.bg.ac.rs

Abstract. This paper explores the application of machine learning in predicting obesity, a significant global health concern. We specifically examine the impact of three feature selection methods — InfoGain, Chi-squared, and ReliefF, on the performance of classification models using Random Forest and Logistic Regression algorithms. By analyzing an obesity dataset categorized into three and seven classes, we identify key features that contribute to model accuracy. The models are evaluated using several metrics: Accuracy, Precision, Recall, Specificity, Sensitivity, and Balanced Accuracy. The findings highlight the role of feature selection in model performance, with the Random Forest algorithm achieving the highest accuracy rate of 96.7%.

Keywords: feature selection, machine learning, classification algorithms, obesity

1 Introduction

The role of machine learning (ML) in healthcare continues to prove its significance and efficacy in various critical areas. By analyzing vast amounts of patient records, lab results, and treatment histories, ML uncovers patterns and trends that otherwise are easy to miss. This process aids in early disease detection, enables personalized treatment plans, improves patient outcomes, and reduces healthcare costs [25]. In this paper, we explore how ML can be used for predicting obesity.

Overweight refers to an excess of fat deposits. Obesity is a chronic and intricate condition characterized by excessive fat accumulation that can negatively impact health. Both are diagnosed by calculating body mass index (BMI) using the formula weight divided by height [28]. The transition from being lean to becoming obese triggers changes in adipose tissue, leads to chronic inflammation and increases the risk of cardiovascular diseases, and contributes to conditions such as stroke. Moreover, obesity is a major factor in insulin resistance, a key element in type 2 diabetes and metabolic syndrome. Additionally, obesity is linked to various cancers including colorectal, pan-

creatic, kidney, and endometrial [8]. According to the [28], obesity among adults worldwide has doubled since 1990, and obesity among adolescents has increased fourfold, resulting in 2.5 billion overweight adults of whom 890 million are living with obesity.

The research objective of our study is to evaluate the effectiveness of feature selection methods Chi-squared, InfoGain, and ReliefF in predicting obesity using Random Forest (RF) and Logistic Regression (LR) algorithms. The research purpose is to demonstrate that these methods can identify relevant features crucial for accurate obesity prediction, leading to high-performance models. This is achieved through experiments in Weka, assessing how feature selection methods impact model performance as dataset complexity increases from 3 to 7 classes.

This paper is structured as follows: in Section 2 we review the literature. Section 3 outlines the dataset, ML algorithms, and feature selection methods that are used. The experimental results and discussion are presented in Section 4 and Section 5 respectively, while Section 6 concludes the paper.

2 Related Work

Feature selection methods such as InfoGain, ChiSquare, and ReliefF have been applied in various domains to identify significant features, such as in cancer data [14], heart disease [21], bank data [23], and network traffic [26]. Moreover, numerous studies have employed Random Forest [5, 7, 19, 26] or Logistic Regression [4, 5, 10, 19, 21, 26] algorithms to construct predictive models, often evaluating their effectiveness through metrics like accuracy, precision, recall, and F1 score. In addition, many studies have utilized the Weka software for model development [1, 10, 14, 15, 18, 21], highlighting its role as a popular tool for machine learning tasks, including feature selection and model building.

[12] reviewed feature selection methods for medical dataset classification, highlighting challenges in balancing feature relevance and computational complexity and stressing the importance of efficient feature selection.

[7] studied obesity in Bangladesh and classified it as low, medium, or high, using 80% of their dataset for training and the rest for testing, comprising 1100 entries from diverse sources. They tested various ML algorithms and found that LR had the highest Accuracy of 97,09% after applying PCA. [19] used R software to evaluate LR and RF algorithms on an imbalanced dataset of obesity risk factors, predicting obesity as a binary classification problem. Resampling techniques were employed, with RF outperforming LR in Precision, Recall, F1 score, and Balanced Accuracy, particularly with imbalanced data.

[15] emphasized feature selection's importance in ML and proposed a new algorithm based on conditional mutual information. The testing was performed in Weka. [18] focused on breast cancer classification Accuracy using feature selection and ML algorithms. Their focus was on wrapper selection methods in Weka, noting increased Accuracy with feature selection for Bayes Network but decreased Accuracy for SVM. [21] used Weka and algorithms Bayes Net, LR, SGD, and KNN with feature selection methods Chi-squared, ReliefF, and Symmetrical uncertainty to predict if the patient has heart

disease. [26] studied Intrusion Detection Systems performance with feature selection methods Chi-squared, InfoGain, and Recursive Feature Elimination coupled with different ML classifiers. Feature selection methods improved model performance across various classifiers, confirming the importance of feature selection methods. However, the study [23] investigated how the GINI index and InfoGain affect classification Accuracy in the Decision Tree classifier algorithm and concluded that irrespective of dataset imbalance, the classification Accuracy remains consistent between models using the GINI index and InfoGain. [14] compared the PCA-IG model with traditional feature selection methods Gain Ratio, ReliefF, and CfsSubset on breast cancer data using Weka. PCA-IG outperformed in Accuracy, Precision, Recall, and training time when it comes to models built with other classifiers.

When it comes to obesity prediction, several studies have explored the effectiveness of different ML algorithms and feature selection methods. [1] explored ML algorithms for classifying childhood obesity in 6-year-old school children in Malaysia using classifiers: Naïve Bayes, Bayes Net, J48, MLP, and SMO. Feature selection methods used are CfsSubsetEvaluator and Consistency in Weka. The study found that feature selection methods with genetic search enhanced accuracy, with J48 achieving the highest Accuracy at 82,72%. Similarly, [10] analyzed obesity risk factors using PRMT in Weka, identifying Naïve Bayes with 99,2% Accuracy as the best classifier for predicting obesity risk based on factors such as age, BMI, and lifestyle. Chi-square was used to select relevant features and the classifiers built were Naïve Bayes, KNN, Kstar, ZeroR, Random Tree, and LR.

Some studies have utilized UCI Machine Learning dataset for obesity-related research. [3] proposed a model that integrates data mining techniques, including Extremely Randomized Trees, Multilayer Perceptron, and XGBoost, implemented in Python to detect and predict obesity levels. The dataset utilized originates from the UCI Machine Learning Repository. Similarly, [5] used the same UCI dataset and applied machine learning algorithms such as LR, RF, Decision Tree, SVM, Gradient Boosting, and Ada Boost. Based on evaluation metrics like accuracy, precision, recall, and F1 score, the results indicated that the Logistic Regression model achieved the highest prediction accuracy. Additionally, [4] employed Decision Trees, Logistic Regression, and KNN for prediction using the same dataset. However, studies [4] and [5] do not employ feature selection methods, whereas [3] utilizes Recursive Feature Elimination as a wrapper-type feature selection algorithm.

In comparison to prior studies, our work shares several similarities and notable differences. Similarities include the utilization of RF and LR machine learning algorithms and datasets similar to those commonly used in obesity prediction studies. Additionally, we adopt feature selection methods akin to those explored in prior works. However, our study stands out in several aspects. Firstly, we examine a dataset with 3 classes, diverging from the predominant focus on binary classification. Furthermore, we extend our investigation to a dataset with 7 classes, offering a perspective on model performance across a broader spectrum of obesity classification. Finally, our work introduces an expanded set of evaluation metrics. While previous research often relies on Accuracy, Precision, and Recall, we incorporated Sensitivity and Specificity metrics and emphasized the use of Balanced Accuracy.

3 Research Methodology

This section is divided into three parts: a description of the dataset, an overview of the machine learning algorithms used, and an explanation of the feature selection methods employed.

3.1 Dataset

The dataset was obtained from the UCI Machine Learning Repository [16], containing 2111 instances and 17 features. Feature descriptions are given in Table 1.

Table 1. Dataset feature description

Feature name	Type	Values	Description
Gender	Nominal	Male, Female	
Age	Numeric	14 – 61	
Height	Numeric	145 – 199	Height in cm
Weight	Numeric	39 – 173	Weight in kg
Family_history	Nominal	Yes, No	
FAVC	Nominal	Yes, No	Frequent caloric food intake
FCVC	Integer	1 – 3	Frequency of vegetables
NCP	Integer	1 – 4	Number of main meals
CAEC	Nominal	No, Sometimes, Frequently, Always	Food between meals
SMOKE	Nominal	Yes, No	
SCC	Nominal	Yes, No	Caloric consumption monitoring
CH2O	Numeric	1 – 3	Daily water intake
FAF	Numeric	0 – 3	Physical activity frequency
TUE	Numeric	0 – 2	Time using technology
CALC	Nominal	No, Sometimes, Frequently, Always	Consumption of alcohol
MTRANS	Nominal	Public Transport, Automobile, Walking, Bike, Motorbike	

The original target variable had seven values: Insufficient_weight, Normal_weight, Overweight_level_I, Overweight_level_II, Obesity_level_I, Obesity_level_II, and Obesity _level_III. Based on these values, a new target variable named Obesity_Risk was created, containing three possible values: Not_obese (Insufficient and Normal_weight), Overweight (Overweight_level_I and Overweight_level_II), and Obese (other values). The dataset was prepared by removing inadequate values, resulting in

1984 instances. The removed instances included unrealistic values for weight or height and values that were inconsistent when considering weight, height, and age together. Among these, there are 895 obese individuals, 546 overweight individuals, and 543 individuals with normal or insufficient weight. The inclusion of Weight and Height as features in the model was guided by recommendations from an internal medicine specialist and a comprehensive review of existing literature. Notably, studies [7, 10] have also incorporated weight in obesity prediction models.

The experiment was conducted in Weka. Based on the previous studies [10, 14], we chose Chi-squared, InfoGain, and ReliefF. These feature selection methods were employed to determine the 5, 8, and 12 most relevant features. Subsequently, classifiers were created using these features and the RF and LR algorithms. Our dataset encompassed 3 classes initially, with subsequent testing performed using the same dataset but expanded to include 7 original classes.

3.2 Machine Learning Algorithms

Random Forest and Logistic Regression are widely used machine learning algorithms for building predictive models across various domains [4, 5, 19, 21, 26], displaying strong predictive capabilities and consistent performance. These algorithms have also proven effective in prior research [7, 10] for predicting obesity. Our decision to use RF and LR was motivated by these studies, leading us to select these two algorithms as the foundation for our analysis.

A Random Forest is an ensemble learning method consisting of decision trees, collectively forming a “forest”. Each decision tree within the forest is constructed using a random subset of features at each node. During the classification phase, each tree provides a vote and the class receiving the majority of votes is selected as the final prediction [9].

Logistic regression is a statistical method for analyzing the relationship between an outcome and multiple explanatory variables. This approach calculates each variable's impact on the odds ratio of the observed event, enabling the examination of how different factors collectively influence the outcome, avoiding the pitfalls of analyzing variables in isolation [22].

3.3 Feature Selection Methods

Feature selection methods belong to a filter category that evaluates the relevance of features based on the inherent properties of the data. They are characterized by their speed, scalability, and independence from particular learning algorithms, requiring selecting features once and then assessing their effectiveness using different classifiers [12]. The selection methods used in this paper are filter methods: InfoGain, Chi-squared, and ReliefF.

To understand InfoGain, it is required to explain entropy. Defined by [20], entropy is a measure of the uncertainty or randomness in a dataset. In classification tasks, entropy quantifies the amount of impurity or disorder in a set of examples. If a dataset

contains instances that belong to different classes, the entropy will be higher. Conversely, if all instances belong to a single class, the entropy will be zero, indicating no uncertainty.

InfoGain is a metric that quantifies the reduction in entropy achieved by splitting the data based on a particular feature. It is used to determine how well a feature separates the data into classes. InfoGain is calculated as the difference between the entropy of the dataset before the split and the weighted sum of the entropies after the split. A higher InfoGain indicates that the feature is more useful for classification as it reduces uncertainty (or entropy) about the target class after the split [23]. In other words, the feature is more informative for classifying the instances.

The Chi-square statistic is a test that measures the degree of association between categorical variables. It evaluates how much the observed data deviates from what would be expected under the null hypothesis, which assumes that the two categorical variables are independent. A high value indicates a significant difference between the observed and expected frequencies, suggesting that there is a strong association between the variables. Low value, on the other hand, suggests that the observed and expected frequencies are close, indicating that the variables are likely independent or have a weak association.

In machine learning, the Chi-square test is commonly used to select the most important features when dealing with categorical data. The Chi-square statistic is computed for each feature by comparing the observed frequencies (actual data) with the expected frequencies (what would be expected if the feature was independent of the class labels). Features are ranked based on their Chi-square values. Features with higher Chi-square values are considered more relevant as they have a stronger correlation with the target variable [13]. This ranking can guide the selection of features to use in model building. Although the Chi-square test is a powerful tool for feature selection, it does have some limitations such as assuming independence of observations, sensitivity to sample size and applicability to categorical data only.

The Relief algorithm is a feature selection method that assesses how well features distinguish nearby instances. The key idea behind Relief is to evaluate the relevance of features by considering their ability to separate instances that are similar (neighbors) but belong to different classes. The algorithm randomly selects an instance from the dataset. The nearest hit is the closest instance to the selected instance that belongs to the same class. The nearest miss is the closest instance to the selected instance that belongs to a different class. For each feature, the algorithm updates a weight based on its ability to distinguish between the selected instance and its nearest hit and miss [11]. If a feature has a similar value for the selected instance and the nearest hit (same class), it is less useful and its weight is decreased. After sampling different instances, the algorithm aggregates the feature weights, ranking them according to their ability to differentiate between instances of different classes.

While the original Relief algorithm is effective, it has several limitations. ReliefF is an extension of the original algorithm designed to address its limitations. Key enhancements in ReliefF are (1) multi-class capability, (2) dealing with missing values, (3) use of multiple neighbors, (4) noise resilience and (5) weight update mechanism [17].

4 Experimental Results

For both datasets, selection methods InfoGain, Chi-squared, and ReliefF were employed using the entire dataset. This approach was chosen because if feature selection is conducted solely on the training set, the selected features or their importance rankings may vary significantly with different random states of the train-test split. This variability can lead to inconsistencies in feature selection, making it difficult to generalize the importance of features. Additionally, evaluating feature importance on the entire dataset provides a more accurate assessment of which features are generally influential. This approach is supported by several studies in literature, including [6, 15, 24].

The experiment was done in Weka, using 10-fold cross-validation, meaning that the dataset is divided into 10 equal-sized subsets. Then, the model is iteratively trained on nine of these subsets and its performance is evaluated on the remaining subset [2]. This method was chosen because it provides a more robust estimate of the model's performance compared to a single train-test split.

The metrics observed are Accuracy, Precision, Recall, Specificity, Sensitivity, and Balanced Accuracy. Accuracy is the measure of correctly classified instances. Precision reflects the accuracy of positive predictions, while Recall (Sensitivity) quantifies the model's ability to identify positive instances. Specificity evaluates the model's capability to correctly identify negative instances. Balanced Accuracy is the mean accuracy considering Sensitivity and Specificity, offering a balanced assessment of model performance, especially in scenarios with imbalanced class distributions [27]. Apart from Accuracy, Precision, and Recall which are commonly observed metrics, we incorporated the Specificity metric to assess the model's ability to differentiate instances not belonging to a specific class. This is crucial in applications where the cost of false positives is high, such as in medical diagnostics. Furthermore, we included Balanced Accuracy, which can be a valuable metric when there is not a similar balance among classes within the dataset. Balanced Accuracy provides a more equitable measure of performance by considering both Sensitivity and Specificity, ensuring that the model is not biased towards the majority class. This is essential for creating robust models that perform well across all classes.

In this section, the findings are discussed first for the 3-class dataset, followed by an analysis of the results obtained from the dataset with 7 classes.

4.1 Results with 3-class Target Variable

Feature ranking based on the method is given in Table 2. It can be noted that Chi-squared and InfoGain give similar rankings, while ReliefF results differ. This difference can be featured in the underlying methodologies used. Chi-square and InfoGain both rely on statistical measures to assess the relevance of features. On the other hand, ReliefF focuses on evaluating feature relevance by considering values between nearest neighbors from the same and different classes.

Table 2. Feature ranking for 3-class dataset

Rank	InfoGain	Chi-squared	ReliefF
1	Weight	Weight	Weight
2	Family_history	CAEC	Family_history
3	CAEC	Family_history	CAEC
4	Age	Age	FCVC
5	NCP	NCP	NCP
6	FAF	FAF	TUE
7	FAVC	FAVC	FAF
8	TUE	TUE	Height
9	FCVC	FCVC	Age
10	SCC	Height	CH2O
11	Height	MTRANS	MTRANS
12	MTRANS	SCC	FAVC
13	CALC	CALC	Gender
14	Gender	Gender	SCC
15	CH2O	CH2O	CALC
16	SMOKE	SMOKE	SMOKE

Weight, Family_history, CAEC, and NCP consistently rank in the top five features across all three methods, meaning these features are highly influential in predicting and understanding obesity in the dataset. Age, FAF, and TUE consistently rank high, but with slight variations in their specific orders across methods. Gender, CALC, and SMOKE consistently rank towards the bottom, suggesting that they have minimal direct impact. MTRANS, FCVC, CH2O, SCC, Height, and FAVC are features whose rankings fluctuate the most across different feature selection methods. This fluctuation in rankings indicates that these features may have varying degrees of influence on predicting or understanding obesity depending on the specific methodology used for feature selection.

The ReliefF feature selection method produces distinct rankings compared to InfoGain and Chi-squared, particularly in the middle and lower ranks. While Weight, Family_history, CAEC, and NCP remain consistently influential, appearing in the top five across all methods, the ReliefF method shows variability with other features. Notably, FCVC and TUE, ranked fourth and sixth by ReliefF, contrast with their more consistent middle rankings by InfoGain and Chi-squared. Features such as Height, FAVC, and SCC exhibit significant ranking shifts under ReliefF, suggesting their influence on predicting obesity varies notably with this method.

Results for models built with InfoGain and Chi-squared feature rankings are given in Table 3. In terms of performance metrics, RF constantly outperforms LR. For 5 and 8 features, RF significantly outperforms LR, indicating that even with fewer features RF can effectively capture the complexities of the dataset better than LR. Both classifiers exhibit improvement as the number of features increases and achieve the best results for 12 features, with RF achieving the highest Accuracy at 96,6%. The Balanced Accuracy values are high, suggesting a well-rounded performance across all classes,

with minimal bias towards any specific class. Models built using Chi-squared feature ranking show the same results. The similarity in performance can be attributed to the small variation in feature rankings between Chi-squared and InfoGain, proving their correlation.

Table 3. RF and LR results using InfoGain and Chi-squared ranking

InfoGain, Chi-squared	Accuracy	Precision	Recall	Specificity	Sensitivity	BA
5 features						
RF	90,4%	90,4%	90,4%	95,4%	90,4%	92,9%
LR	83,6%	84,2%	83,6%	91,9%	83,6%	87,8%
8 features						
RF	91,6%	91,6%	91,6%	96%	91,6%	93,8%
LR	85%	85,6%	85%	92,6%	85%	88,8%
12 features						
RF	96,6%	96,6%	96,6%	98,5%	96,6%	97,6%
LR	96,2%	96,2%	96,2%	98,3%	96,2%	97,3%

Results for models built using ReliefF ranking are given in Table 4. Models demonstrate overall higher performance, but also more variations compared to results obtained using InfoGain and ChiSquare feature ranking. RF still achieves higher results than LR. For both algorithms increasing the number of features from 5 to 8 leads to significant improvements in model performance. Both models give the best results with 8 features used, with RF achieving the highest Accuracy at 96,7%. Eight features that demonstrate the best performance are: Weight, Family_history, CAEC, FCVC, NCP, TUE, FAF, and Height. The Specificity metric values, exceeding 92%, indicate the model's proficiency in minimizing false positives, while the high Balanced Accuracy suggests the model's ability to make precise predictions across all classes.

Table 4. RF and LR results using ReliefF ranking

ReliefF	Accuracy	Precision	Recall	Specificity	Sensitivity	BA
5 features						
RF	86,9%	86,9%	86,9%	93,3%	86,9%	90,1%
LR	83,8%	84,4%	83,8%	92,1%	83,8%	88%
8 features						
RF	96,7%	96,7%	96,7%	98,5%	96,7%	97,6%
LR	96,6%	96,6%	96,6%	98,5%	96,6%	97,6%
12 features						
RF	96,6%	96,6%	96,6%	98,5%	96,6%	97,6%
LR	95,8%	95,8%	95,8%	98,2%	95,8%	97%

4.2 Results with 7-class Target Variable

In this section, we show the model performance variations when using a 7-class dataset. The evaluation metrics used are Accuracy and Recall, as we primarily focus on feature selection rather than assessing overall model performance. Feature ranking based on the method is given in Table 5. Again, Chi-squared and InfoGain give similar rankings, while ReliefF results differ.

Table 5. Features ranking for 7-class dataset

Rank	InfoGain	Chi-squared	ReliefF
1	Weight	Weight	Gender
2	Age	Age	Weight
3	FCVC	FCVC	FCVC
4	Gender	CAEC	Family_history
5	CAEC	Gender	CAEC
6	Family_history	Family_history	CALC
7	NCP	Height	MTRANS
8	Height	NCP	NCP
9	CALC	FAF	Height
10	FAF	CALC	TUE
11	MTRANS	MTRANS	Age
12	TUE	TUE	FAF
13	FAVC	FAVC	FAVC
14	SCC	SCC	CH2O
15	CH2O	CH2O	SCC
16	SMOKE	SMOKE	SMOKE

Models constructed using InfoGain and Chi-squared consistently produce similar results, despite differences in feature ranking. Notably, when selecting subsets of 5, 8, and 12 features, both methods identify the same groups of features. RF still achieves higher results than LR. For both algorithms, increasing the number of features from 5 to 8 leads to significant improvements, with RF achieving the highest Accuracy at 94% with 8 features. However, Accuracy decreases going from 8 to 12 features for both classifiers (Table 6).

Table 6. RF and LR results using InfoGain and Chi-squared ranking, 7-class dataset

InfoGain, Chi-squared	Accuracy	Recall
5 features		
RF	85,4%	85,4%
LR	73,8%	73,8%

8 features		
RF	94%	94%
LR	91,5%	91,5%
12 features		
RF	93,6%	93,6%
LR	91,3%	91,3%

Using ReliefF for feature selection, there is a more pronounced improvement in Accuracy and Recall as the number of features increases (Table 7). Both RF and LR models exhibit significant enhancements in performance from 5 to 8 features and further improvements at 12 features. The best Accuracy at 93,6% is achieved with RF, using 12 features.

Table 7. RF and LR results using ReliefF ranking, 7-class dataset

ReliefF	Accuracy	Recall
5 features		
RF	79%	79%
LR	73,9%	73,9%
8 features		
RF	86,5%	86,5%
LR	75,1%	75,1%
12 features		
RF	93,6%	93,6%
LR	91,3%	91,3%

5 Discussion

In both datasets, Weight, Age, CAEC, and Family_history maintain relatively high rankings across different methods, indicating their importance. Features FCVC, Gender, MTRANS, CALC, and Height become significantly more important in the 7-class dataset compared to the dataset with 3 classes, while TUE, FAF, and FAVC lose relevance. Features SMOKE and CH2O consistently rank towards the bottom, suggesting that they have a minimal direct impact, regardless of the dataset. RF consistently outperforms LR. For the 3-class dataset, the highest accuracy of 96.7% is achieved using the ReliefF method with RF and 8 features. In the case of the 7-class dataset, RF achieves the highest accuracy of 94% using 8 features with InfoGain or Chi-squared rankings.

In a 3-class dataset, the InfoGain and Chi-squared feature selection methods demonstrate a trend of increasing accuracy as more features are added, with the best results achieved at 12 features. However, the ReliefF method exhibits a different pattern: classifiers reach their peak performance with just 8 features. Adding more features beyond this point does not improve performance, suggesting that a more streamlined feature subset may be more effective for model optimization. In contrast, in a 7-class dataset,

the patterns shift. When using InfoGain and Chi-squared for feature selection, classifiers achieve the highest accuracy with 8 attributes. However, classifiers built using ReliefF continue to improve as more features are added, peaking at 12 features.

This variation highlights how the performance of feature selection methods is influenced by the number of classes in the dataset. While InfoGain and Chi-squared methods show optimal performance at 12 features in a 3-class dataset, they perform best with 8 features in a 7-class dataset. Conversely, ReliefF, which peaks at 8 features in a 3-class dataset, reaches its highest accuracy with 12 features in a 7-class dataset. These results show that both the choice of feature selection method and the optimal number of features are closely tied to the dataset's class structure.

The study [7] presents relevant findings for comparison with our work.

Table 8. Comparison with study in [7]

ML algorithm	Research	Accuracy	Precision	Recall
LR	Our study	96,6%	96,6%	96,6%
	[7]	97,09%	97%	97%
RF	Our study	96,7%	96,7%	96,7%
	[7]	72,3%	57%	72%

The authors of [7] worked with a dataset comprising 1100 entries and 28 features to classify obesity into 3 classes: low, medium, or high. They employed ML algorithms KNN, SVM, LR, Naïve Bayes, RF, Decision Tree, Ada Boosting, MLP, and Gradient Boosting. Comparing overall performance, their study achieved a slightly higher Accuracy rate of 97,09% employing LR and PCA. In contrast, our study involves a larger dataset with 1984 entries and 16 features. We used feature selection methods instead of PCA. These methods focus on selecting a subset of features based on their relevance, whereas PCA transforms the entire set of features into new variables. Comparing algorithm performance, the authors of [7] achieved better results using LR. Our work demonstrated significantly better performance with RF, particularly in terms of Accuracy and Precision. Notably, we incorporated Balanced Accuracy as an evaluation metric, and our study achieved a rate of 97,6% using RF and 8 features selected with ReliefF.

5.1 Potential Limitations

Despite our thorough analysis, there are several potential limitations to this research that should be acknowledged. One concern is the specificity of the dataset, as the findings may not generalize well to other datasets with different characteristics. We did not include validation using datasets from other sources, which would demonstrate the generalizability and robustness of the results beyond the dataset used in this study. Furthermore, the focus on the effectiveness of Chi-squared, InfoGain, and ReliefF for feature selection, without including results from models that do not use these methods, may limit understanding of the full impact of feature selection on model performance. The

study could also benefit from the inclusion of additional feature selection methods and machine learning models, which we plan to explore in future research.

6 Conclusion

Our study demonstrated that feature selection methods can effectively reduce the number of model features while maintaining comparable performance in classification models, especially in the context of obesity prediction. Through experimentation in Weka, we have identified several key features, including Weight, Age, FCVC, CAEC, and Family_history. Notable findings highlight the superiority of RF over LR and the best model built with RF having an Accuracy rate of 96,7%. The transition from a 3-class to a 7-class dataset emphasizes the increased significance of the feature selection method for the Accuracy metric. InfoGain and Chi-square methods maintain consistent and reliable feature rankings and groupings, showcasing their suitability for this purpose. ReliefF exhibited variability in feature rankings compared to InfoGain and Chi-squared, contributing to noticeable improvements in model performance as the number of selected features increased. This variability highlights ReliefF's efficacy in identifying relevant features that contribute to enhancing model accuracy and robustness. Effective feature selection greatly enhances model performance overall.

In the future, our research aims to expand into data preprocessing techniques to enhance input data quality, address class imbalance issues, and incorporate additional feature selection methods. We will continue to refine our models and explore their applicability to more extensive datasets. Additionally, some more advanced tools could be utilized for creating and testing models, which would allow for further analysis on how these tools could enhance the results and insights of the study.

Acknowledgments. This paper is funded by the Faculty of Organizational Sciences, University of Belgrade.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdullah, F. S., Manan, N. S., Ahmad, A., Wafa, S. W., Shahril, M. R., Zulaily, N., . . . Ahmed, A. (2016). Data Mining Techniques for Classification of Childhood Obesity Among Year 6 School Children. *Recent Advances on Soft Computing and Data Mining* (pp. 465–474). Bandung: Springer. doi:https://doi.org/10.1007/978-3-319-51281-5_47
2. Berrar, D. (2019). Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology* (Vol. 1, pp. 542-545). Elsevier. doi:10.1016/B978-0-12-809633-8.20349-X
3. Choudhuri, A. (2022, January). A hybrid machine learning model for estimation of obesity levels. In International Conference on Data Management, Analytics & Innovation (pp. 315-329). Singapore: Springer Nature Singapore.

4. Cui, T., Chen, Y., Wang, J., Deng, H., & Huang, Y. (2021, May). Estimation of Obesity levels based on Decision trees. In 2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM) (pp. 160-165). IEEE.
5. Devi, K. N., Krishnamoorthy, N., Jayanthi, P., Karthi, S., Karthik, T., & Kiranbharath, K. (2022, January). Machine Learning Based Adult Obesity Prediction. In 2022 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-5). IEEE.
6. Elemam, T., & Elshrkawey, M. (2022). A highly discriminative hybrid feature selection algorithm for cancer diagnosis. The Scientific World Journal, 2022(1), 1056490.
7. Ferdowsy, F., Rahi, K. A., Ismail , J., & Habib, T. (2021). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences, 2*. doi:https://doi.org/10.1016/j.crbeha.2021.100053
8. Fruh, S. M. (2017). Obesity: Risk factors, complications, and strategies for sustainable long-term weight management. *Journal of the American Association of Nurse Practitioners*, S3-S14. doi:10.1002/2327-6924.12510
9. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier Inc. doi:https://doi.org/10.1016/C2009-0-61819-5
10. Hossain, R., Mahmud, H., Hossin, A., Noori, S., & Jahan, H. (2018). PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques. *International Conference on Computational Intelligence and Data Science Proceedings*, *132*, pp. 1068-1076. doi:https://doi.org/10.1016/j.procs.2018.05.022
11. Kira, K., & Rendell, L. A. (1992). The feature selection problem: Traditional methods and new algorithm. *AAAI-92: Tenth National Conference on Artificial Intelligence Proceedings* (pp. 129-134). San Jose: AAAI. doi:https://dl.acm.org/doi/10.5555/1867135.1867155
12. Mwadulo, M. W. (2016). A Review on Feature Selection Methods For Classification Tasks. *International Journal of Computer Applications Technology and Research, 5*(6), 395-402. doi:10.7753/IJCATR0506.1013
13. Na, S., An, X., Yan, C., & Ji, S. (2020). Incremental Feature Reduction Method Based on Chi-Square Statistics and Information Entropy. *IEEE Access, 8*, 98234-98243. doi:10.1109/ACCESS.2020.2997013
14. Omuya, E. O., Okeyo, G. O., & Kimwele, M. W. (2021). Feature Selection for Classification using Principal Component Analysis and Information Gain. *Expert Systems with Applications, 174*. doi:https://doi.org/10.1016/j.eswa.2021.114765
15. Phyu, T. Z., & Oo, N. N. (2016). Performance Comparison of Feature Selection Methods. *MATEC Web of Conferences.* doi:10.1051/matecconf/20164206002
16. Repository, U. (2019). *Estimation of Obesity Levels Based On Eating Habits and Physical Condition.* Retrieved April 2024, from UC Irvine Machine Learning Repository: https://doi.org/10.1016/j.dib.2019.104344
17. Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning, 53*, 23-69. doi:https://doi.org/10.1023/A:1025667309714
18. Saoud, H., Ghadi, A., Mohamed, G., & Abdelhakim, B. A. (2019). Using Feature Selection Techniques to Improve the Accuracy of Breast Cancer Classification. *The Proceedings of the Third International Conference on Smart City Applications*, (pp. 307-315). doi:10.1007/978-3-030-11196-0_28
19. Sewpaul, R., Awe, O. O., Dogbey, D. M., Sekgala, M. D., & Dukhi, N. (2023). Classification of Obesity among South African Female Adolescents: Comparative Analysis of Logistic Regression and Random Forest Algorithms. *International Journal of Environmental Research and Public Health*, *21*(1), 2.
20. Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, 27*(3), 379-423. doi:https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

21. Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digital Health*. doi:10.1177/2055207620914777
22. Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 12-18. doi:10.11613/BM.2014.003
23. Tangirala, S. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm. *(IJACSA) International Journal of Advanced Computer Science and Applications, 11*(2), 612-619.
24. Tariq, M. A. (2024). A Study on Comparative Analysis of Feature Selection Algorithms for Students Grades Prediction. Journal of Information and Organizational Sciences, 48(1). https://doi.org/10.31341/jios.48.1.7
25. Tekieh, M. H., & Raahemi, B. (2015). Importance of Data Mining in Healthcare: A Survey. *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1057-1062). Paris: IEEE. doi:http://dx.doi.org/10.1145/2808797.2809367
26. Thakkar, A., & Lohiya, R. (2021). Attack classification using feature selection techniques: a comparative study. *Journal of Ambient Intelligence and Humanized Computing*, *12*(1), 1249-1266.
27. Vujovic, Z. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications Volume, 12*(6), 599-606. doi:10.14569/IJACSA.2021.0120670
28. WHO. (2024, March 1). *Obesity and overweight*. Retrieved April 14, 2024, from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight