Journal of Gono Bishwabidyalay
Vol.4 Issue:1
January 2023

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

PERFORMANCE ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS TO PREDICT OBESITY

Farzana Tasnim [1], Shatabdee Bala[1], Umme Honey[1], Tania Akter[1], and Md. Karam Newaz[1]*

Abstract

Obesity is a medical condition characterized by a fast rise in body fat. Obesity is becoming increasingly common. The worldwide obesity rate has proved that this is a serious health issue over the last decade. We collected data from specific urban and rural areas containing various risk factors of our daily food habits such as fast food per week, cold drinks per week, tea with sugar, vegetable per week, and physical records such as TV shows per day, sleep per day, daytime sleep, daily physical exercise, and so on. We were able to get 819 data points based on 18 variables. Then, according to the BMI score, we labeled the class of each record in the data set as normal, overweight, and obese. Our data set contains a total of 249 normal, 203 overweight, and 367 obese data. It includes 235 females and 584 males ages between 10 to 82. Weka is used to preprocess the data, which includes deleting duplicate instances and partial data, fixing missing values, and double-checking the data. And also, to determine the accuracy and error measurement of a model to examine the obesity risk factor class. We employ seven widely used classification methods, including BN, NB, HT, LR, MP, NB Tree, and RS. We use metrics such as accuracy, kappa, precision, recall, and f-measure to assess their performance. The findings show that the Bayes Net method classification model works better than other research with the same antecedents, having the highest weighted value with 98.78% accuracy, 99.7% ROC, 98.82% precision, 98.82% recall, and 98.82% f-measure. The analysis of the results focuses on controlling this cardiovascular disease in Bangladeshi rural and urban people.

Keywords: Classification, Machine learning Algorithm, Performance, Obesity, Overweight

1. Introduction

Obesity refers to a state of excessive fat accumulation throughout the body. There are several factors, including diet and environmental ones that contribute to obesity. Today, obesity is a major health concern all over the world. People are progressively adopting an unhealthy lifestyle, as seen by their consumption of excessive junk food, late-night sleeping, and prolonged sitting. In particular, adolescents are affected by their unconscious attitudes. Malignancies, diabetes,

[1]Dept. of CSE, Gono Bishwabidyalay, Savar, Dhaka-1344, Bangladesh.
Corresponding Author: email: rekha.tasnim@gmail.com, Phone: +8801765524232.

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

metabolic syndrome, and cardiovascular disease are only a few of the many chronic diseases that are made worse by the prevalence of overweight and obesity as lifestyle conditions (Safaei, 2021,p. 01). And now, obesity is the fifth biggest cause of mortality worldwide. By 2030, the World Health Organization (WHO, 2021) predicts that lifestyle-related illnesses would account for 30% of all deaths worldwide.

Obesity has roughly tripled worldwide since 1975. More than 1.9 billion adults aged 18 and over were overweight in 2016. Over 650 million of them were fat. In 2016, 39% of adults aged 18 and over were overweight, while 13% were obese. Bangladesh, a low-income South Asian country, has seen an increase in the prevalence of overweight and obesity. Bangladesh, like many low- and middle-income countries, has seen demographic and nutritional changes in its population, such as changes in lifestyle (e.g., high-calorie food intake, sedentary lifestyle) and urbanization. In Bangladesh, the prevalence of significant chronic health disorders has progressively increased, with deaths from chronic diseases increasing from 8% in 1986 to 68% in 2006 (Chowdhury, 2015, p. 01).

Researchers have worked hard to find out factors that influence the generation of obesity, even developing web tools like the calculation of body mass index (BMI) (World Health Organization, «Body mass index calculator,» WHO), where one can determine a person's level of obesity. However, these tools only calculate BMI and don't take into account other important factors, like whether the person has a family history of obesity, how much time they spend exercising, or their gene expression profiles. So, we need a smart tool that can find the risk of obesity quickly and accurately.

Information and communication technologies (ICTs) are moving in this direction, especially artificial intelligence (AI) and machine learning (ML). ML techniques are now very important for the early diagnosis of diseases like diabetes (Shahriare, 2020, p. 453), cesarean section (Rahman, 2021, p. 293), blood cancer (Tasnim, 2022, p. 09), heart disease (Satu, 2018, p. 01) etc.

This study will analyze obesity datasets using a variety of machine learning algorithms to create better results for early obesity detection that are more accurate. In particular, the ML models are investigated in this work to indicate numerous risk factors for the rising incidence of obesity among persons living in urban and rural areas of Bangladesh. This article contributes

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

- To collect an obese dataset with random responders.
- To evaluate the performance of different machine learning algorithms on this obese dataset.
-To find out the best classification algorithm based on evaluation matrix accuracy, roc, precision, recall, and f-measure.

This paper is organized as follows: section 2 shows previous research and similar approaches, section 3 Materials and Methods describes the dataset and methods used to generate the machine learning model, section 4 describes the experimental results and discussions, and section 5 presents the conclusion.

2. Literature Review

In this section, we will describe several similar works that inspired us to work on obesity risk prediction using local data. We also mentioned some earlier research in which machine learning and data analysis were utilized to forecast the risk level of obesity or employed different machine learning approaches for classifying diabetes datasets using different machine learning tools.

To combat this global epidemic, (Alqahtani, 2021, p. 103) used supervised data mining techniques such as Random Forest and Multi-Layer Perception (MLP) to build an effective way of tracking obesity levels. The data came from 14-61-year-olds in nations including Mexico, Peru, and Colombia, with diverse eating habits and physical conditions. The data set contains 2111 instances and 17 attributes labeled with Obesity, which allows for data categorization using Overweight Levels I and II, Insufficient Weight, Normal Weight, and Obesity Type I through III. This study discovered that the Random Forest method outperformed the MLP algorithm in terms of accuracy, with an overall classification rate of 96.7 percent.

(Pang, 2021, p. 02) examined seven machine learning models created to predict childhood obesity from the age of 2 to 7 years using EHR data up to the age of 2 years. After removing implausible growth data with severe quality control, 27,203 (50.78% male) patients remained for model development. To predict obesity, seven machine learning models were built. Multiple conventional classifier metrics were used to evaluate model performance, and the differences between seven models were analyzed using Cochran's Q test and post hoc pairwise testing. XGBoost outscored all other models with an AUC of 0.81 (0.001). It outperformed all other models statistically significantly on basic classifier measures (sensitivity set at 80%): precision

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

30.90% (0.22%), F1-score 44.60% (0.26%), accuracy 66.14% (0.41%), and specificity 63.27 % (0.41 %).

(Hossain, 2018, p. 1069) collected 259 data points from various urban and rural areas about various risk factors associated with our everyday activities and simulated the risk factors using statistical tools (SPSS), which can assist in forecasting the major risk factor of obesity by comparing the class level attribute to other attributes in a cross-sectional study. The outcome of this process is age (0.002), height (0.002), weight 0.000), healthy lifestyle (0.000), marital status (0.001), BMI (0.000), economic (0.028), sleep per day (0.011) has a significant link with our obesity class, according to the P-value ($p<0.05$). This work suggested a risk mining technique (PRMT) that predicts a model to analyze the risk factor of an obesity class utilizing various data mining classifiers with WEKA used to estimate the accuracy and error measurement. As a result of this method, Naive Bayes is the best classifier for the 10-fold cross-validation research. Thamrin et al.

(Thamrin, 2021, p. 02) evaluated the ability of ML methods, specifically Logistic Regression, and Regression Trees (CART), and Naive Bayes, to detect the presence of obesity using available public health data, to use a novel approach with sophisticated ML methods to predict obesity in an attempt to go beyond traditional prediction models, and to compare the performance of three different methods. Furthermore, they resolved data imbalance by predicting obesity status based on risk factors in the dataset using the Synthetic Minority Oversampling Technique (SMOTE). According to the findings of this investigation, the Logistic Regression approach performs the best. Despite this, kappa coefficients indicate only moderate agreement between expected and measured obesity. Location, marital status, age groups, education, sweet drinks, fatty/oily foods, grilled foods, preserved foods, seasoning powders, sugary drinks, alcohol, mental-emotional issues, diagnosed hypertension, physical exercise, smoking, and fruit and vegetable consumption all play a role in adult obesity.

To identify obesity levels based on lifestyle, (Santisteban Quiroz, 2022, p. 02) built a computational intelligence model using supervised and unsupervised data mining techniques such as the Light Gradient Boosting Machine (Light GBM) classifier, random forest (RF), decision tree (DT), Extremely Randomized Trees (ET), and logistic regression (LR). The primary source of data for this study was a survey of 2,111 people aged 14 to 61 from Colombia, Mexico, and Peru. The study collected data on the main causes of obesity, intending to point to

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

high-calorie consumption, decreased energy expenditure owing to lack of physical exercise, eating disorders, heredity, and socioeconomic variables. The results displayed that the Light GBM classification model has an accuracy of (0.9745) and a weighted value of AUC (0.9990). Ferdowsy et al.

(Ferdowsy, 2021, p. 02) moved towards a machine-learning-based pathway for predicting the risk of obesity using different machine-learning algorithms. They collected more than 1100 data from many varieties of people of different ages and collect information from both suffering obesity and non-obesity. For this research, they applied nine prominent machine learning algorithms. They employed k-nearest neighbor, random forest, logistic regression, multilayer perceptron (MLP), support vector machine (SVM), nave Bayes, adaptive boosting (ADA boosting), decision tree, and gradient boosting classifiers. Experimental results show high, medium, and low obesity. Logistic Regression Algorithm has the greatest accuracy at 97.09%.

Hammond et al. (Hammond, 2019, p. 02) used electronic health records and publicly available data to predict childhood obesity. They developed several machine learning techniques for binary classification and regression. By gathering data from the first two years, they demonstrated that they could anticipate with reasonable accuracy. It forecasts that children would be obese by the age of five. They used logistic regression, a random forest, and a gradient boosting model to predict their dichotomous measurements of low obese/medium obese/high obese, and LASSO regression to predict their continuous BMI values. They repeated the bootstrap 100 times to obtain the best model performance for the outcome.

The authors (Molina, 2021, p. 2526) developed software for estimating obesity based on the SEMMA data mining methodology, employing machine learning techniques such as Decision Trees (DT), Bayesian Networks, and Logistic Regression (LR). The data used were the results of a study conducted in Colombia, Mexico, and Peru, in which 712 university students aged 18 to 25 took part. A survey was used to understand the behavior of obese people and identify the level of obesity through questions based on physical features, social factors, and others. Following model training, three metrics were used to validate the models: The authors did not examine the ROC curve in terms of recall, true positive rate, and false-positive rate. Because of its high precision, the decision tree model was chosen as the optimal strategy (97.4%). H. Rossman et al. (Rossman, 2021, p. 133) presented an intelligent algorithm for predicting the high risk of childhood obesity before the greatest increase in BMI occurs. From 2002 to 2018, 132.262 computerized medical health records of Israeli children were used. Diagnoses, prescribed medications, child and family data, lab tests, and demographic information were all included in

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

the data. The models were trained using data from the first two years of life, and the risk of obesity was calculated between the ages of five and six. "The gradient boosting trees technique was trained on a portion of the data, and the model's quality was tested by calculating the area under the ROC curve and the area under the Precision-Recovery curve."

3. Materials and Methods

In this section, we will go over the many methods and materials that were utilized in the course of this research.

3.1 Data

We collected factors from several publications and journals, and we also spoke with dietitians and nutrition specialists to obtain the primary factors of obesity. We collected data both online and offline. We gathered the information we needed from a variety of sources. Offline data collection refers to club members, and school and university students in Dhaka and nearby cities. On the other hand, collecting data from social media and creating a Google form to gather data from multiple people by sending a link is referred to as online data collection. We gave them the following surveys and recorded their responses on the survey paper. We gathered all of the relevant information based on this.

3.2 Dataset Description

We were able to get 819 data points based on 18 variables. Then, according to the BMI score, we labeled the class of each record in the data set as normal, overweight, and obese. Our data set contains a total of 249 normal, 203 overweight, and 367 obese data. It includes 235 females and 584 males ages between 10 to 82. Fig. 1 depicts the distribution of the records.

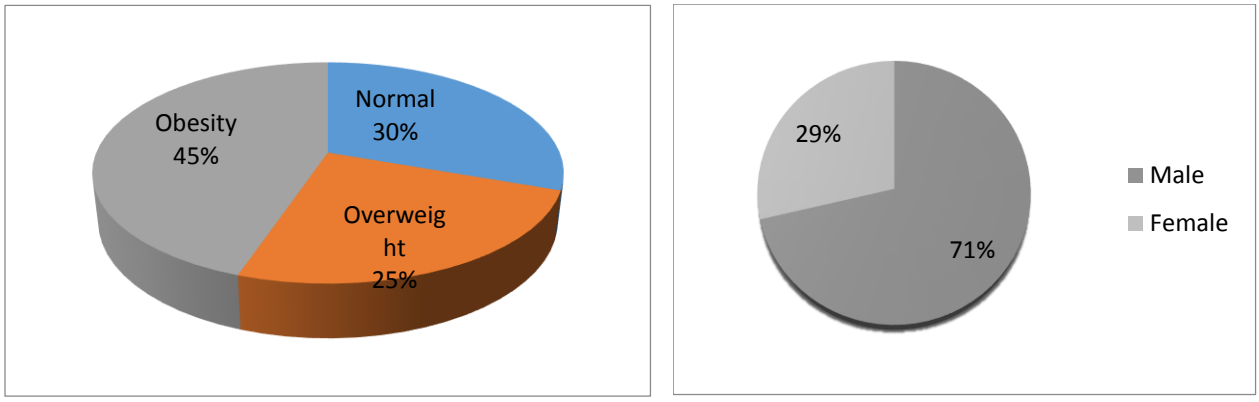

Fig. 1: Distribution of data groups

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

Table1: Dataset description

The attributes defined by the dataset are shown in Table 1.

S.no	Factors	Type	Values
1	Age	Numeric	10-82
2	Gender	Nominal	Male, Female
3	Height	Numeric	Height of the person in meter
4	Weight	Numeric	Weight of the person in kg
5	Marital Status	Nominal	Married, Unmarried
6	Education	Nominal	None, SSC, HSC, Graduated
7	Employment	Nominal	Yes, No
8	Household Income per Month	Numeric	Total family income in a month
9	Physical Exercise	Nominal	Yes, No
10	Diabetes	Nominal	Yes, No
11	Use TV or Internet Per Day	Numeric	Total hours of using TV or internet
12	Take Tobacco	Nominal	Yes, No
13	Residence	Nominal	City, Village
14	Sleeping Hour Per Day	Numeric	Total Hours of sleep in 24 hours
15	Healthy LifeStyle	Nominal	Yes, No
16	Any Medication	Nominal	Yes, No
17	BMI	Numeric	Body mass index (calculation of height and weight)
18	Type	Nominal	Normal, Overweight, Obese

3.3 Methods

Fig. 2 depicts the technique and workflow for the analysis. The diagram systematically describes the procedures, tasks, and approaches.

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

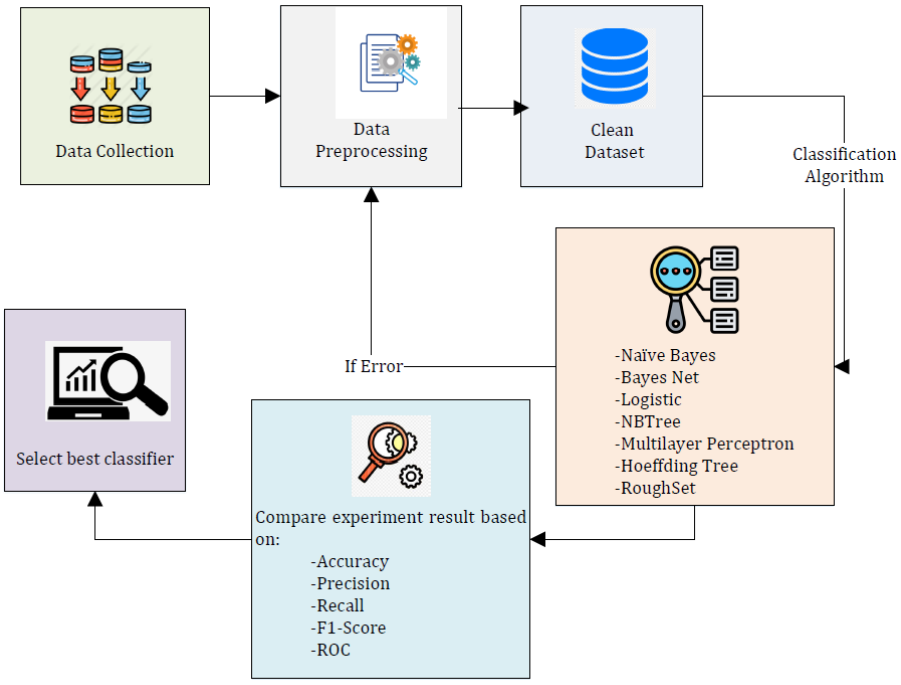

Fig. 2: Workflow of the research methodology

3.3.1 Data Pre-processing

Weka software is used to preprocess the data, which includes deleting duplicate instances and partial data, fixing missing values, and double-checking the data. For replacing values in this raw dataset, the Replacing Missing Values (semi-supervised technique) is utilized. After completing the data preprocessing, we apply the correlation feature selection (CFS) assessment method in conjunction with the best-first search (BFS) to extract the best relevant characteristics for this study.

3.3.2 Classification algorithms

Bayes Net: Bayes Net is combined with a directed acyclic graph and a set of conditional probability tables (Williams, 2006, p. 10). A node represents features/classes and a link represents the relation between them. A conditional probability table is used to determine the strength of the links. If a node has no parents, then the probability distribution is unconditional. Besides, if a node has one or more parents, the probability of each parent depends on their parents. In the network, there are two distinct strategies for estimating the conditional probability table. Search is done by K2, TAN, hill-climbing, simulated annealing, tabu search, and genetic algorithm.

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

Naive Bayes: Na¨ıve Bayes is implemented the probabilistic Na¨ıve Bayes classifier. It is analyzed the relationship between attribute and class for each instance to derive conditional probability. Naive Bayes can use kernel density estimators which are improved when the normality assumption is incorrect and can use numeric attributes using supervised discretization (Witten & Frank, 1999, p. 371).

Heoffding Tree: The Heoffding Tree is a tree structure like a decision tree that is used for the data stream. It is performed well in high dimensionality where the node is expanded to split enough statistical evidence. The decision tree is not worked well to classify documents of high-dimensional feature vectors (Domingos, 2000, p. 74).

Logistic regression: Many classification methods generate probability rather than binary classifications. The obvious example is Naive Bayes, but other methods do as well. The values obtained by linear regression in the preceding lesson may appear to be probabilities, but they are not. However, a version known as "logistic regression" generates probabilities. Logistic regression is a powerful classification that uses the "logit transform" to predict probabilities directly (Zhang, 2009, p. 452).

Multilayer Perceptrons: Multilayer Perceptrons are perceptron networks or linear classifier networks. In reality, they can use "hidden layers" to construct arbitrary decision limits. Weka features a graphical interface that allows you to design your network structure with as many perceptrons and links as you want. A brief test revealed that a multilayer perceptron with one hidden layer outperformed other approaches on two of six data sets - not bad. However, it was 10–2000 times slower than other approaches, which was a disadvantage. It's a classifier that learns a multi-layer perceptron to classify cases using backpropagation. The network can be constructed by hand or using a simple algorithm. During training, the network settings can also be monitored and changed. This network's nodes are all sigmoid (except for when the class is numeric, in which case the output nodes become un thresholded linear units) (Gardner & Dorling, 1998, p. 2628).

NB Tree: Kohavi developed the NB Tree method in 1996 to improve the accuracy of the Naive Bayes algorithm. NB Tree, which generates a mix of decision-tree and Naive-Bayes classifiers: the decision-tree nodes include univariate splits as standard decision-trees, but the leaves contain Naive-Bayesian classifiers. The method preserves the interpretability of Naive-Bayes and

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

decision trees while producing classifiers that commonly outperform both constituents, particularly in bigger databases evaluated (Kohavi, 1996, p. 205)**.**

Rough Set: The rough set is a mathematical method for dealing with partial and unclear data. This theory serves as a foundation for converting data into information. Rough set theory is used to detect unnecessary and redundant data, as well as data dependencies between large volumes of data. This raw set is utilized in a variety of data mining-related applications (Chan, 1998, p. 172).

3.3.4 Evaluation metrics

To measure the performance of various classifiers, several assessment metrics such as Accuracy, F1-score, precision, and recall are used (Cook & Ramadas, 2020, p. 145) (Akter et al., 2021, p. 745). These measurements, however, are altered through the use of true positive (TP), true negative (TN), false positive (FP), and false-negative (FN), which can be characterized as follows :

-Accuracy is defined as the ratio of correct forecasts to total predictions, as shown below.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad \text{----------(1)}$$

-F1-Score collects precision and recall and manipulates their harmonic mean.

$$\text{F-Measure} = 2 * \frac{precision * recall}{precision+recall}$$

$$= \frac{TP}{TP+\frac{1}{2}(FN+FP)} \quad \text{----------(2)}$$

-Precision is a metric that assesses how many of the items that were anticipated to be positive cases are.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{---------- (3)}$$

-The recall is the percentage of correctly identified affirmative cases.

$$\text{Recall} = \frac{TP}{P} \quad \text{----------(4)}$$

4 Results and Discussions

4.1 Results

All of the experiments in this research are done in WEKA (Waikato Environment for Knowledge Analysis) version 3.8. Validation is performed on numerous data and properties on an obese database, including those that have repeated bounces, similar values, and missing values. For replacing values in this raw dataset, the Replacing Missing Values (semi-supervised technique)

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

is utilized. Then we used seven commonly used classification algorithms, including BN, NB, HT, LR, MP, NB Tree, and RS. First and foremost, we justify several classifiers algorithm with 5 folds cross-validation method. Several metrics were used to assess their performance, including accuracy, roc, precision, recall, and f-measure. Table 2 displays the classification results of various classifiers.

Table 2: classification results of various classifiers

Classifier	Accuracy	ROC	Precision	Recall	F1-Score	Type
Bayes Net	98.78	1	1	1	1	Normal
		0.995	0.985	0.966	0.975	Over Weight
		0.997	0.981	0.992	0.986	Obese
Weighted Average		0.997	0.988	0.988	0.988	
Naïve Bayes	90.96	0.99	0.949	0.892	0.919	Normal
		0.975	0.912	0.813	0.859	Over Weight
		0.987	0.886	0.975	0.929	Obese
Weighted Average		0.985	0.911	0.91	0.909	
Hoeffding Tree	90.11	0.981	0.915	0.904	0.909	Normal
		0.929	0.825	0.768	0.796	Over Weight
		0.991	0.93	0.973	0.951	Obese
Weighted Average		0.973	0.899	0.901	0.9	
Logistic Regression	95.73	0.999	0.98	0.98	0.98	Normal
		0.991	0.924	0.901	0.913	Over Weight
		0.995	0.96	0.973	0.966	Obese
Weighted Average		0.995	0.957	0.957	0.957	
Multilayer Perceptron	90.11	0.994	0.935	0.928	0.931	Normal
		0.95	0.793	0.813	0.803	Over Weight
		0.989	0.94	0.932	0.936	Obese
Weighted Average		0.981	0.902	0.901	0.901	
NB Tree	97.44	0.998	0.98	0.984	0.982	Normal
		0.988	0.965	0.956	0.96	Over Weight
		0.995	0.976	0.978	0.977	Obese
Weighted Average		0.994	0.974	0.974	0.974	
Rough Set	91.70	0.951	1	0.996	0.998	Normal
		0.892	0.99	0.961	0.975	Over Weight
		0.935	0.979	0.997	0.988	Obese
Weighted Average		0.929	0.988	0.988	0.988	

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

All of the classifiers provided results that were over 90% accurate. The Bayes Net method, on the other hand, produces the best results, with 98.78 percent accuracy, 99.7% ROC, 98.8 percent precision, 98.8 percent recall, and 98.8 percent f-measure. Then, in order, NB Tree, LR, RS, NB, HT, and MP produce better results.

4.2 Performance analysis of classifier

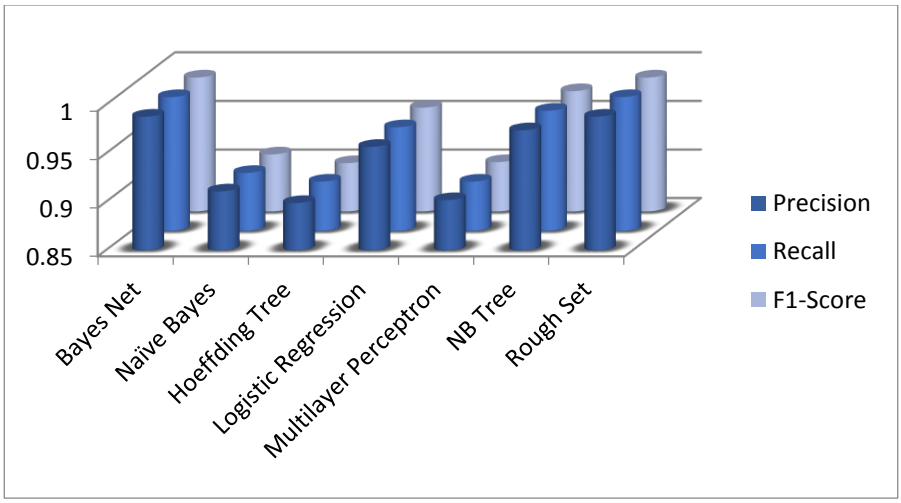

Fig. 3: Performance comparison among classification algorithms

The results shown in Fig.3 demonstrated that the Bayes Net and Rough Set algorithms offer the highest levels of precision, recall, and F1-Score, which are all 0.988 respectively. Following the previous graph is another one that displays the results of the average accuracy level.

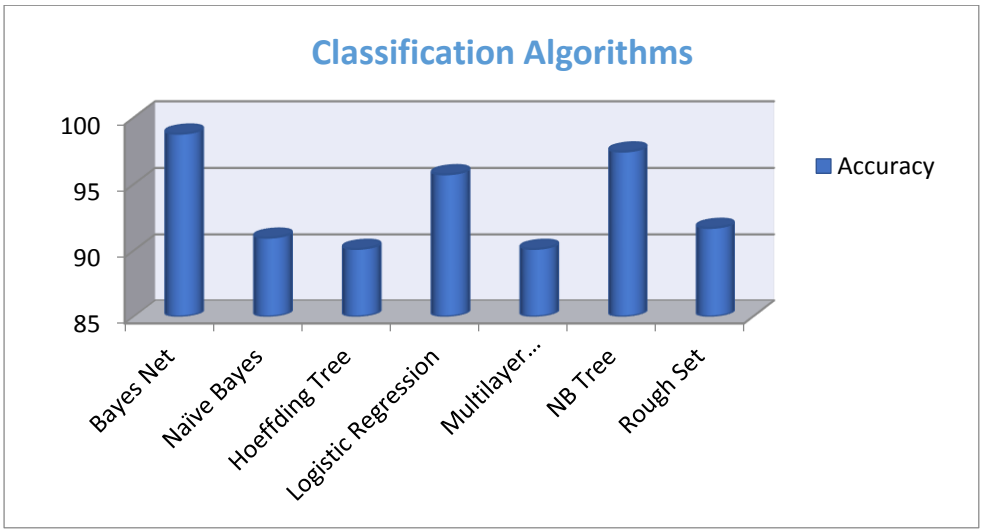

Fig. 4: Accuracy level of different classification algorithms

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

Based on this chart given in Fig.4, we can conclude that of the several classification algorithms, Bayes Net is the one that provides the most accurate results which are 98.87%. Where the Naive Bayes and logistic models perform less well on the accuracy scale from the list of algorithms. The accuracy of the others is around average. Therefore, Bayes Net is the optimal classifier for this endeavor.

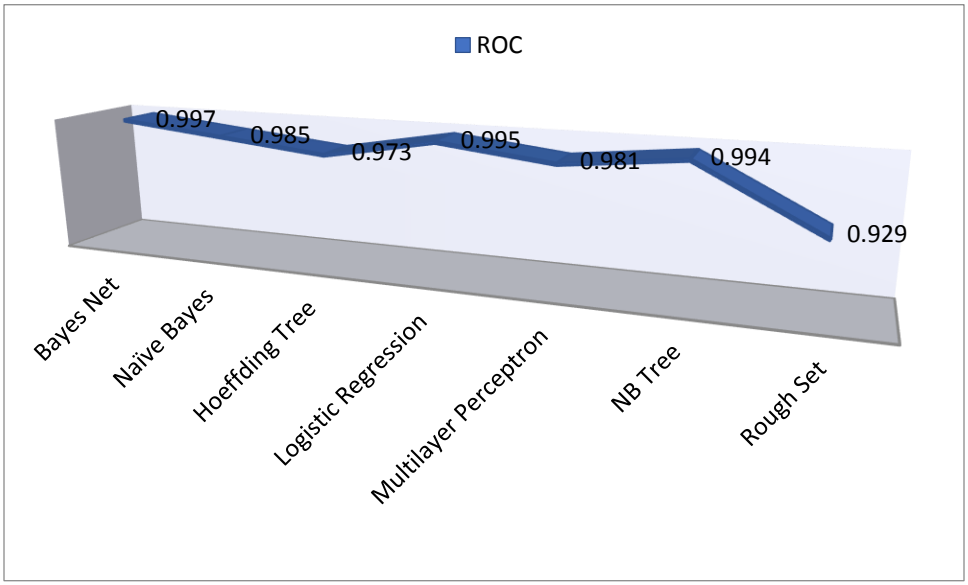

Fig. 5: Receiver Operating Characteristic Curve

The roc area was derived through seven algorithms (Fig. 5). Where it is clear that the ROC area of Bayes Net (99.7%) is the best. This demonstrates that Bayes Net has the best performance.

5 Conclusion

We used many different algorithms in our work, including the Bayes Net, naive Bayes, Hoeffding Tree, Logistic, Multilayer perceptron, NB Tree, and Rough Set on a dataset including information about a variety of people from Bangladesh. To determine which classifier algorithm is more effective, we compare their outputs. The Bayes Net method performed the best, with an accuracy of 98.72%, and the ROC curve value for Bayes Net algorithm is also high at 99.7%. Based on our results, Bayes Net is the optimal tool for tackling obesity-related datasets. Data is an increasingly significant component of research; in the future, the dataset will be expanded, and some deep learning models will be utilized to predict obesity.

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

References

Akter, T., Khan, M. I., Ali, M. H., Satu, M. S., Uddin, M. J., & Moni, M. A. (2021). Improved Machine Learning based Classification Model for Early Autism Detection. *International Conference on Robotics, Electrical and Signal Processing Techniques*, 742–747. https://doi.org/10.1109/ICREST51555.2021.9331013

Alqahtani, A., Albuainin, F., Alrayes, R., Al Muhanna, N., Alyahyan, E., & Aldahasi, E. (2021). Obesity Level Prediction Based on Data Mining Techniques. *IJCSNS International Journal of Computer Science and Network Security*, *21*(3), 103. https://doi.org/10.22937/IJCSNS.2021.21.3.14

Chan, C. C. (1998). A rough set approach to attribute generalization in data mining. *Information Sciences*, *107*(1–4), 169–176. https://doi.org/10.1016/S0020-0255(97)10047-0

Chowdhury, M. A. B., Uddin, M. J., Khan, H. M. R., & Haque, M. R. (2015). Type 2 diabetes and its correlates among adults in Bangladesh: a population-based study. *BMC Public Health*, *15*(1), 1–11. https://doi.org/10.1186/s12889-015-2413-y

Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. *Stata Journal*, *20*(1), 131–148. https://doi.org/10.1177/1536867X20909693

Domingos, P. (n.d.). *Mining High-Speed Data Streams*. 71–80.

Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, *2*(May), 100053. https://doi.org/10.1016/j.crbeha.2021.100053

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmospheric Environment*, *32*(14–15), 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0

Hossain, R., Mahmud, S. M. H., Hossin, M. A., Haider Noori, S. R., & Jahan, H. (2018). PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques. *Procedia Computer Science*, *132*, 1068–1076. https://doi.org/10.1016/j.procs.2018.05.022

Hammond, R., Athanasiadou, R., Curado, S., Aphinyanaphongs, Y., Abrams, C., Messito, M.J., Gross, R., Katzow, M., Jay, M., Razavian, N. and Elbel, B., 2019. Predicting childhood obesity using electronic health records and publicly available data. PloS one, 14(4), p.e0215571.

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, *7*(Utgo 1988), 202–207. http://www.aaai.org/Library/KDD/1996/kdd96-033.php

Molina Estren, D., De la Hoz Manotas, A.K. and Mendoza Palechor, F., (2021). Classification and features selection method for obesity level prediction.

Pang, X., Forrest, C. B., Lê-Scherban, F., & Masino, A. J. (2021). Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics*, *150*(March). https://doi.org/10.1016/j.ijmedinf.2021.104454

Rahman, S., Khan, M. I., Satu, M. S., & Abedin, M. Z. (2021). Risk prediction with machine learning in cesarean section: Optimizing healthcare operational decisions. *Intelligent Systems Reference Library*, *192*(October), 293–314. https://doi.org/10.1007/978-3-030-54932-9_13

Rossman, H., Shilo, S., Barbash-Hazan, S., Artzi, N.S., Hadar, E., Balicer, R.D., Feldman, B., Wiznitzer, A. and Segal, E., 2021. Prediction of childhood obesity from nationwide health records. The Journal of Pediatrics, 233, pp.132-140.

Safaei, M., Sundararajan, E. A., Driss, M., Boulila, W., & Shapi'i, A. (2021). A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Computers in Biology and Medicine*, *136*(August), 104754. https://doi.org/10.1016/j.compbiomed.2021.104754

Santisteban Quiroz, J. P. (2022). Estimation of obesity levels based on dietary habits and condition physical using computational intelligence. *Informatics in Medicine Unlocked*, *29*(July 2021), 100901. https://doi.org/10.1016/j.imu.2022.100901

Satu, M. S., Tasnim, F., Akter, T., & Halder, S. (2018). Exploring Significant Heart Disease Factors based on Semi-Supervised Learning Algorithms. *International Conference on Computer, Communication, Chemical, Material and Electronic Engineering, IC4ME2 2018*. https://doi.org/10.1109/IC4ME2.2018.8465642

Shahriare Satu, M., Atik, S. T., & Moni, M. A. (2020). *A Novel Hybrid Machine Learning Model to Predict Diabetes Mellitus*. *September*, 453–465. https://doi.org/10.1007/978-981-15-3607-6_36

ISSN: 2521-828X
2706-9303 (Online)
2706-9311(CD-ROM)

Tasnim, F., Ahmed, T., Ahmed, K., Patwary, M. F. K., & Newaz, M. K. (2022). Potential Risk Factors and Association of Significant Factors of Blood Cancer in Bangladesh Using Data Mining Techniques. *Journal of Engineering Science*, *13*(1), 9–20. https://doi.org/10.3329/jes.v13i1.60558

Thamrin, S. A., Arsyad, D. S., Kuswanto, H., Lawi, A., & Nasir, S. (2021). Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018. *Frontiers in Nutrition*, *8*(June), 1–15. https://doi.org/10.3389/fnut.2021.669155

Williams, N., Zander, S., & Armitage, G. (2006). Preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *Computer Communication Review*, *36*(5), 7–15. https://doi.org/10.1145/1163593.1163596

Witten, I. H., & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (The Morgan Kaufmann Series in Data Management Systems)*. *31*, 371. http://www.amazon.com/Data-Mining-Techniques-Implementations-Management/dp/1558605525

Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I., & Keane, J. (2009). Comparing data mining methods with logistic regression in childhood obesity prediction. *Information Systems Frontiers*, *11*(4), 449–460. https://doi.org/10.1007/s10796-009-9157-0