BMI Prediction Using Machine Learning (ML) Methodologies

Supraja Bathula (700735187)
Machine Learning
sxb51870@ucmo.edu

Jagadeesh Naidu Nelakurthi(700730988)
Machine Learning
jxn09880@ucmo.edu

Bhanu Prakash Dhusari(700726879)
Machine Learning
bxd68790@ucmo.edu
(AND) Vikram Reddy Settypalli
(700740663)
Machine Learning
vxs06630@ucmo.edu

Abstract— Obesity and overweight are complex, multifaceted, and global public health concern that affects individuals of all ages and increases their risk for a variety of illnesses. This score is used to quantify obesity and overweight, as well as identify those at risk for obesity and overweight. BMI is calculated by dividing the individual's weight in kilograms by the square of his height in meters (kg/m2). According to WHO statistics, a body mass index (BMI) of less than 25 is considered normal, whereas a BMI of more than 25 poses a risk for overweight and obesity. Approximately 38 percent of the global senior population will be obese by 2030. It is anticipated that by 2030, up to 57.8% of the world's older population would suffer from obesity or overweight. Obesity and overweights are rising quickly in both developed and developing nations. This research aimed to evaluate the classification of BMI using many machine learning techniques.

Keywords:- BMI, Obesity, SVM, GcForest, random forest

I. INTRODUCTION

According to the World Health Organization (WHO), overweight and obesity are defined as the excessive accumulation of fat in different sections of the body. It is regarded as a significant public health issue since it is associated with a variety of illnesses and even death. Among the disorders linked to obesity include type 2 diabetes mellitus, hypertension, stroke, osteoarthritis, depression, Alzheimer's, and many forms of cancer, including breast, prostate, kidney, ovary, liver, and colon cancer. In this context, an adult is overweight if his Body Mass Index (BMI) is more than 25 kg/m2 and obese if it is greater than 30 kg/m2. In addition, according to WHO analyses, the global prevalence of obesity has almost quadrupled over the years, becoming a concern not just in affluent nations but also in emerging nations.

Europe and America had the greatest prevalence of overweight in 1975, with respective values of 40.0% and 35.0%; these values climbed to 62.3% and 63.3% in 2016, respectively. Concerning obesity, the worldwide prevalence in 1975 was 4.3%, whereas the prevalence in 2016 was 13.2%. Similarly, the areas with the greatest incidence of obesity in 1975 were Europe and the United States, with values of 10.1% and 9.0%, while the prevalence of overweight climbed to 25.3% and 29.0%, respectively, in 2016.

Alternatively, if the rising trend of obesity prevalence continues, it is anticipated that by 2030 almost half of the world's population would be overweight or obese. Some studies assert that the increase in the global prevalence of obesity and overweight is attributable to complex changes in the population, including changes in lifestyle, increased calorie consumption, decreased physical activity, and other factors such as urbanization, environmental changes, socioeconomic status, and genetic changes. Therefore, preventing obesity is difficult because it requires changes in the population's physical activity and eating habits, as well as collective support from the government, industry, and scientific and medical communities, in order to minimize overweight by empowering the population to make sensible lifestyle decisions.

In this regard, the literature has several studies proposing the creation of technological tools, such as the use of machine learning, to assist professionals in the field with decision-making in order to minimize the incidence of obesity and overweight. Machine Learning is the study of computing techniques for identifying complicated patterns in millions of data to construct prediction models. In recent years, machine-learning approaches in the health sector have gained popularity. The purpose of this article is to construct a prediction model utilizing machine learning methods and survey data in order to identify individuals with obesity and overweight and make timely judgments.

Numerous factors have been identified as risk factors for obesity; in general, they fall into the following categories: demographic and socioeconomic (gender, age, education, income, marital status, and urban areas), lifestyle (fast food consumption, stress, smoking, alcoholic beverage consumption, and low levels of physical activity), and genetic (genetic factors) (obese parents). Some of these risk factors can be addressed or altered, but others cannot. Implementing a successful risk reduction plan urgently calls for the identification of modifiable risk factors for obesity at the individual and community level. Numerous research have investigated more effective methods for anticipating obesity using the information at hand. Machine Learning (ML), which is now one of the most well-liked subjects in the scientific community for large-scale datasets, is used in a unique approach that was only just presented to provide an answer to this issue.

Utilizing ML techniques to model epidemiological data is becoming more and more common. These techniques may enhance our knowledge of general health, including the spread of diseases, their detection, the identification of risk factors for health issues, and therefore, chances for intervention. Numerous ML techniques and algorithms have been used to analyze obesity-related health data among other things. In the case of obesity, it is crucial to create a precise data categorization to make it easier to identify predicted risk variables from the available data in an attempt to manage these risk factors and finally lower obesity-related morbidity and mortality.

All of the remaining pieces of paper have been divided into six sections for your convenience. Introduction: In Section I, we provide a high-level overview of the subject matter. Providing an overview of the research methodology is covered in Section 2. This section also contains a more in-depth discussion of machine learning-based classifiers, their performances, and the datasets that they use. The findings are presented in Section III. Section IV presents a synopsis of our findings and conclusions as a result of our investigation. Additionally, the subject of future work is covered in this section.

II. RELATED WORK

The American government classified childhood obesity as an epidemic in 1998. The U.S. Surgeon General published a call to action in 2001 to promote certain activities with relation to this public health concern. [9] In response to this appeal, Arkansas launched a programme for BMI monitoring and screening in the state's public schools (Act 1220 of 2003. HB 1583). Act 1220 of 2003 was the first piece of law mandating public schools to calculate each student's BMI and send parents and guardians private letters known as "Child Health Reports." [10,11] Since then, several additional states have put some variant of this law into place via screening or monitoring systems.

Programs for monitoring obesity concentrate on overall rates of obesity and calculate the proportion of kids in a school or district that are underweight, at a healthy weight, overweight, or obese. Parents who participate in screening programmes might learn about their child's BMI category. Since then, at least 25 states have approved variants of this law mandating public schools to calculate pupils' BMI, and several of these states further demand public schools to provide parents/guardians health reports. [12,13]

By increasing parental knowledge of their children's unhealthful weight status and encouraging parenting behaviours that support a healthy body weight, such BMI screening programmes might lower childhood obesity. However, there is conflicting information about how beneficial screening programmes are. Almond et alanalysis .'s of the effects of overweight reports on students enrolled in New York City public schools employed a regression discontinuity methodology. They were able to acquire accurate but modest estimations showing that kids who were classified as overweight in one period were not significantly more likely to have a lower BMI or weight the next year than kids who were classified as normal weight. Prina and Royer[15] carried out an experimental screening programme in Mexico and found that BMI reports successfully informed parents about obesity but did not significantly change their behaviour. A growing body of data that questions the effectiveness of accurate judgments of weight status and future weight increase may be one rationale for null results. [16–18] Self-perception of being overweight is linked to higher future weight gain in adolescents and young adults, while the underlying processes remain unknown. [16,17] Similarly, weight gain has been seen in younger children whose parents thought they were overweight. [18]

The usefulness of BMI screening programmes is a problem overall. Concerns about the novelty of the information supplied and whether BMI was a reliable indicator of a healthy weight were expressed by parent focus groups created to examine the content of parental notification letters from the Massachusetts programme. [19] Unintended outcomes including stigmatisation and body dissatisfaction have been mentioned as causes for worry. [9] However, one research that looked at weight-based bullying in Arkansas before and two years after school-based BMI screening was implemented showed no rises in bullying overall or among overweight or obese teenagers. [20] More work is required in this area.

The informative utility of BMI screening programmes is one topic that has not been addressed in the literature to yet. These provide longitudinal data that aid in a better comprehension of childhood obesity, its causes, and the efficiency of initiatives to encourage a healthy body weight in children. [21] An significant feature that was overlooked in early critiques of BMI screening programmes is the ability to more accurately identify those who are at risk of becoming obese in the future. [9,22,23] From early infancy until preadolescence, the prevalence of obesity tends to rise. [1] By more effectively addressing the children who have a high possibility of becoming fat, obesity prevention efforts may be amplified if the availability of BMI information early in elementary school enhances the capacity to identify children who are at highest risk of becoming obese. Until now, this benefit of early BMI screening programmes in public schools may have gone unnoticed.

The goal of this research is to evaluate the educational benefit of the first and longest-running BMI screening programme in the country, the Arkansas programme. No other research has looked at the possible educational usefulness of BMI screening programmes to identify kids at risk of obesity.

In order to determine the significance of BMI data during kindergarten (usually aged 5–6 years) on identifying children who are most likely to be obese by the fourth grade (often aged 9–10 years), this research uses a number of machine learning methods. This research specifically seeks to determine if having access to BMI data in kindergarten enhances predictions of obesity in fourth grade beyond those that would otherwise be possible to make in the absence of the screening programme.

III.   METHODOLOGY

For making our predictive model analysis we will need to proceed with certain steps

- Dataset loading & Preparation
- Data visualization
- Assign our data instances and the target value (X and y columns)
- Split data into training and test sets
- Train our model (Deep Forest & SVM & RandomForest)
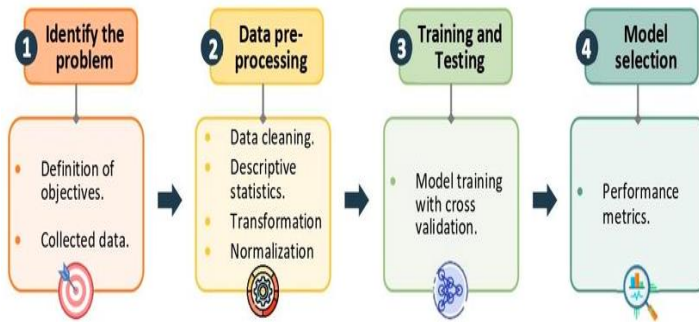- Test and evaluate the model

Fig 1a. Pipeline diagram

A. Datasets

The data we use here is the set of 500 with columns Gender, Height, weight, and index. The index has 5 values. They are:
0= Extremely weak: BMI<16
1=Weak: 16<BMI<18.5
2=Normal: 18.5<BMI<24.9
3= Overweight:25<BMI<29.9
4=Obesity:30<BMI 34.9
5= Extreme obesity: BMI>35

The dataset determines how accurate the algorithms are, thus we cleanse the dataset first before improving results. Perform feature extraction if indeed the information is enormous. Age in years, weight in kilos, height in centimeters, sexuality like a male or female, BMI throughout weight per square meters, and 0 to 6 rates of obesity - are among the factors inside the database having respective measurements specified.

Whenever one or even more classes control the overall entire set of information as core categories and perhaps other categories were sporadic or subordinate courses, it creates a data imbalance. Although sparsity will provide a bad categorization predictive performance for the small category, it will generate a high categorization predictive performance for the main type.

B. Data Preparation

1. Dataset balancing

Classification models trained with imbalanced data tend to produce biased predictions with incorrect results, therefore in many circumstances, classes with fewer occurrences are insufficient for the model, and a sub-process is required to balance the data [21, 22]. In this instance, the classes in the gathered dataset were uneven; hence, the "oversampling" approach was used to balance the minority classes in the training dataset.
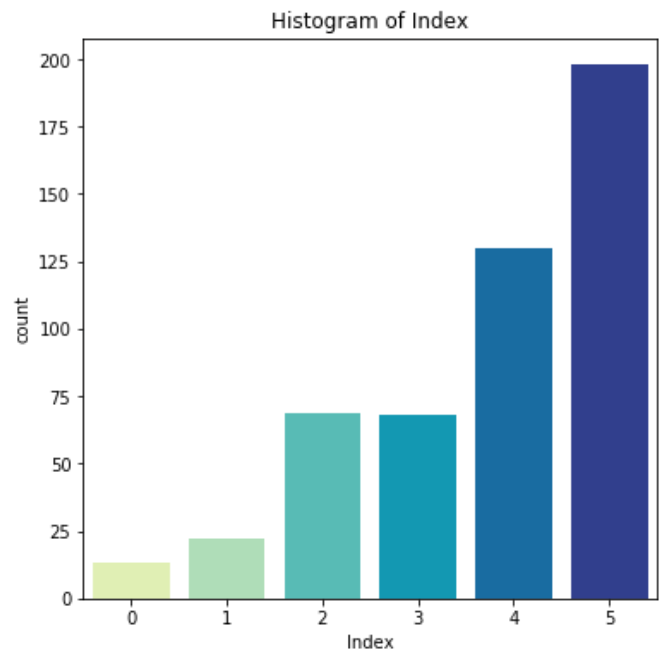
2. Categorical data encoding

About 80% of the variables in the utilized database were categorical, necessitating the employment of data transformation methods since certain machine learning algorithms prohibit the inclusion of non-numerical data. The attributes of the gender property are encoded using ordinal encoding rather than one hot encoding. To change the target feature's classes, the label encoding approach was used.

C. Data Visualization

Index Histogram

A histogram is a bar graph-like display of data that groups a variety of categories into horizontal columns along the x-axis. The vertical y-axis displays the frequency or percentage of occurrences in each column of data. Columns may be used to display data distribution patterns.

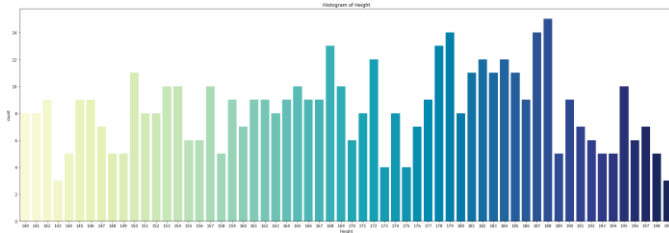Fig 2a. Frequency of values falling under each Index

[0,1,2,3,4,5]

Fig 2b. Frequency of values falling under certain height
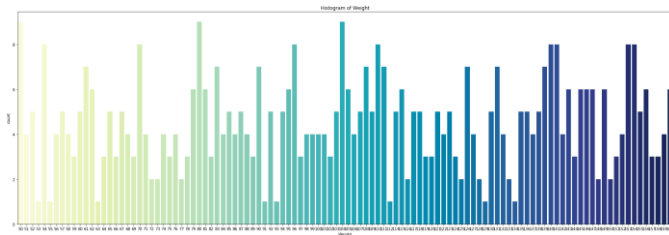
intervals

Fig 2c. Frequency of values falling under certain weight

intervals

Pair plot

To plot multiple pairwise bivariate distributions in a dataset, you can use the pair plot() function. This shows the relationship for (n, 2) combination of variables in a DataFrame as a matrix of plots, and the diagonal plots are the univariate plots.
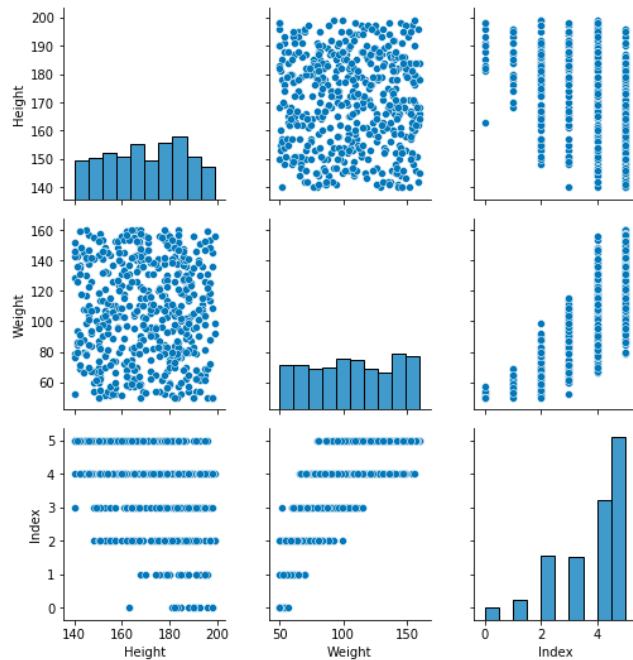

Fig 3. Pair plot

Correlations

A scatterplot displays the strength, direction, and form of the relationship between two quantitative variables. A correlation coefficient measures the strength of that relationship. Calculating a Pearson correlation coefficient requires the assumption that the relationship between the two variables is linear.
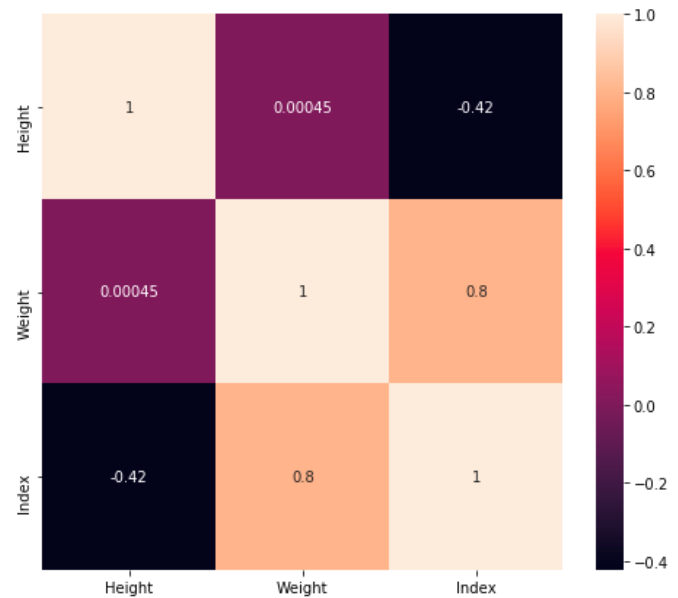

Fig 4. Correlation

D. Modeling

Considering the original dataset, we designate the Index column as the 'target' variable y and the rest columns as 'predictors' for variable X. To further separate them, put arbitrary integers such as 1 to Female and 0 to Male (equivalent to a Boolean/True value) in the Gender column.

Many machine learning techniques need scaling to bring all features to the same level so that the size of a single important value does not influence the model.

First, I divide data into training and test sets. Then, using the Standard Scaler technique, which assumes that the data is normally distributed within each feature and scales it so that the distribution is centered at 0 with a standard deviation of 1, I scale the data so that the distribution is centered at 0 with a standard deviation of 1.

I. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm frequently utilized in pattern recognition and

classification challenges for multiclass classification. Python's multiclass support is handled according to a one-vs-one approach, where each classifier distinguishes between two distinct classes, and the combination of all one-vs-one classifiers results in a multiclass classifier.

II. DeepForest/GC forest

DeepForest/GC forest algorithm. Because of this, gcForest uses a cascade structure, in which each level of the cascade receives feature information that has already been processed by the level before it. An ensemble of decision tree forests (also known as an ensemble of ensembles) is used to construct each level of the hierarchy. For the sake of diversity, we've included a variety of forest kinds here. Each forest can produce an estimate of class distribution for a given instance by counting the percentage of different classes of training examples at the leaf node where the instance in question is located, and then averaging the results across all trees in the same forest.

III. Random Forest

Random Forest for Multiclass because it gives the greatest prediction rate in healthcare datasets when compared to other classification algorithms. Random Forest utilizes averaging to increase prediction accuracy and control over-fitting can handle a high number of features, and is useful for predicting which variables are significant in the underlying data being modeled.

E. Validation Method

The models are assessed under the assumption that validation data (V) have the same distribution as training data (T). In the literature, there are a number of assessment metrics; however, a standard metric for measuring the accuracy of prediction models has not yet been created. This paper evaluates the performance of the tested models based on a comparison of several measures. The lines below outline the performance measures used in this study:

Confusion matrix. It helps to evaluate the quality of the classification model and is represented by a matrix, which allows visualizing the performance of each class of the prediction model

Accuracy. Defined as the sum of the total of true positives and true negatives divided by the total number of results

Precision. Defined as the ratio of the number of correctly predicted true positives to the total number of predicted positives.

Recall. It is also known as the true positives rate. Defined as the ratio of the number of correct positives to the total predictions classified as positives

F1-score. A single measure that combines the sensitivity and precision value of the prediction model

IV. RESULTS

In the training and evaluation process, three machine learning methods were tested. The prediction of the tested models was internally validated by the test dataset.

	precision	recall	f1-score	support
1	0.80	0.80	0.80	5
2	0.89	0.89	0.89	9
3	0.71	0.83	0.77	6
4	0.78	0.86	0.82	21
5	0.97	0.88	0.92	34
accuracy			0.87	75
macro avg	0.83	0.85	0.84	75
weighted avg	0.87	0.87	0.87	75

Accuracy Score 0.8666666666666667
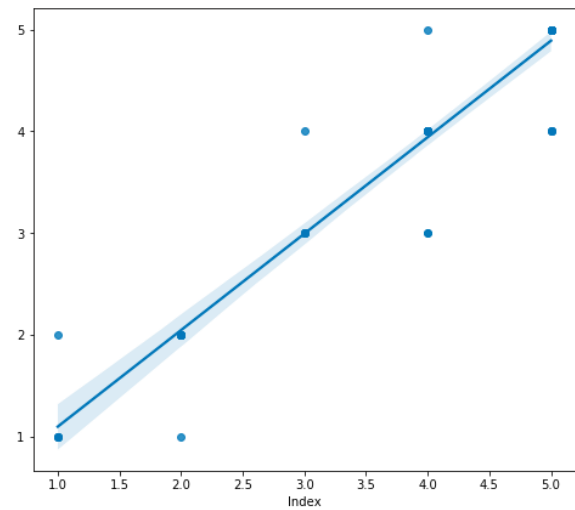Error Rate 0.13333333333333333

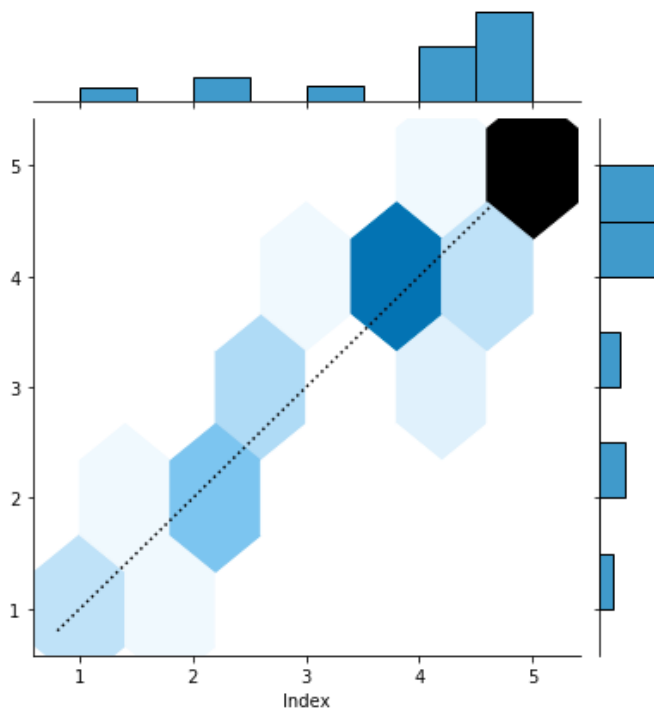Fig 5a. RandomForest Classification Report

Fig 5b. RandomForest Prediction

Fig 5c. RandomForest results

Fig 6c. GcForest Prediction

```
[2022-11-05 05:41:38.193] Start to fit the model:
[2022-11-05 05:41:38.195] Fitting cascade layer = 0
[2022-11-05 05:41:38.658] layer = 0  | Val Acc = 87.059 % | Elapsed = 0.463 s
[2022-11-05 05:41:38.662] Fitting cascade layer = 1
[2022-11-05 05:41:39.145] layer = 1  | Val Acc = 86.588 % | Elapsed = 0.483 s
[2022-11-05 05:41:39.145] Early stopping counter: 1 out of 2
[2022-11-05 05:41:39.147] Fitting cascade layer = 2
[2022-11-05 05:41:39.691] layer = 2  | Val Acc = 85.176 % | Elapsed = 0.544 s
[2022-11-05 05:41:39.691] Early stopping counter: 2 out of 2
[2022-11-05 05:41:39.691] Handling early stopping
[2022-11-05 05:41:39.693] The optimal number of layers: 1
```

Fig 6a. GcForest Training Report

```
[2022-11-05 05:44:22.495] Start to evalute the model:
[2022-11-05 05:44:22.498] Evaluating cascade layer = 0
              precision    recall  f1-score   support

           1       0.80      0.80      0.80         5
           2       0.89      0.89      0.89         9
           3       0.83      0.83      0.83         6
           4       0.79      0.90      0.84        21
           5       0.97      0.88      0.92        34

    accuracy                           0.88        75
   macro avg       0.86      0.86      0.86        75
weighted avg       0.89      0.88      0.88        75

Accuracy Score 0.88
Error Rate 0.12
```
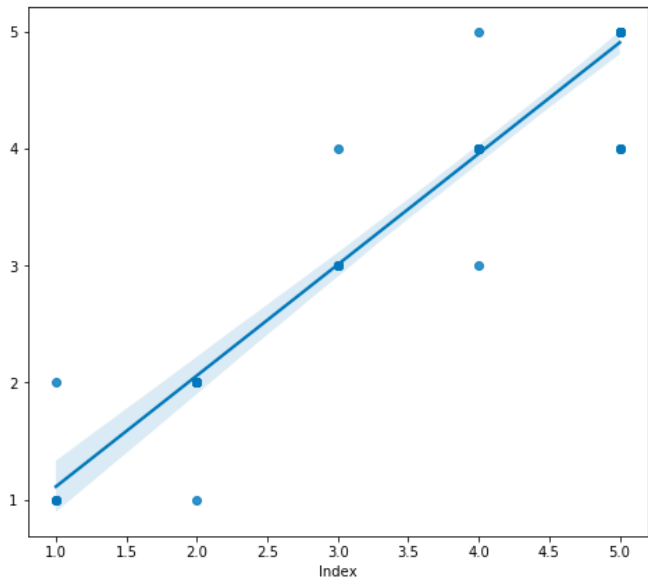
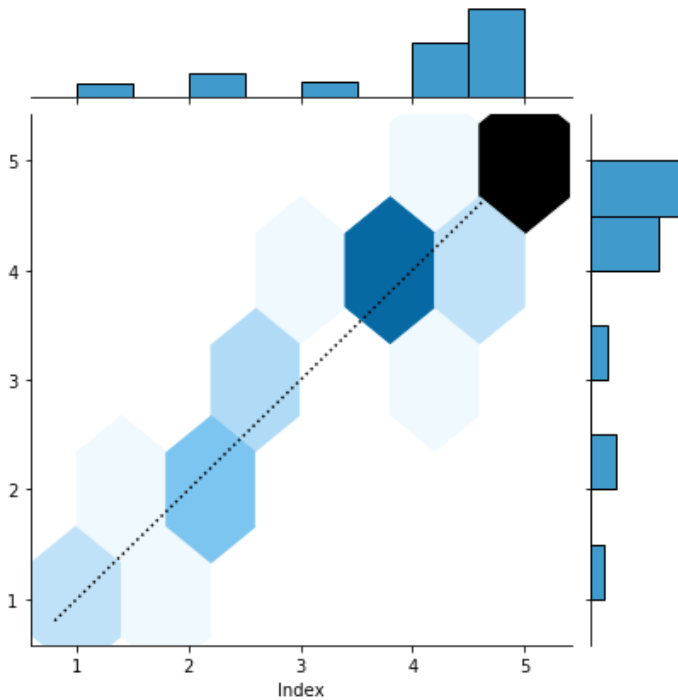Fig 6b. GcForest Classification Report

Fig 6d. GcForest Result

	precision	recall	f1-score	support
1	0.80	0.80	0.80	5
2	0.89	0.89	0.89	9
3	0.56	0.83	0.67	6
4	0.85	0.81	0.83	21
5	1.00	0.94	0.97	34
accuracy			0.88	75
macro avg	0.82	0.85	0.83	75
weighted avg	0.90	0.88	0.89	75

```
Accuracy Score 0.88
Error Rate 0.12
```
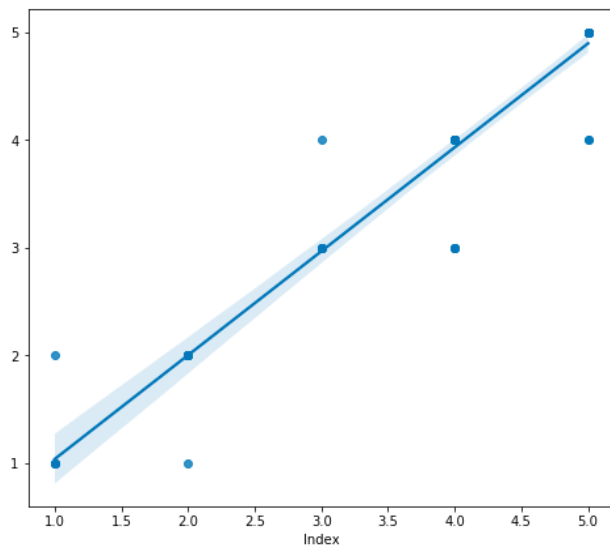
Fig 7a. SVM Classification Report
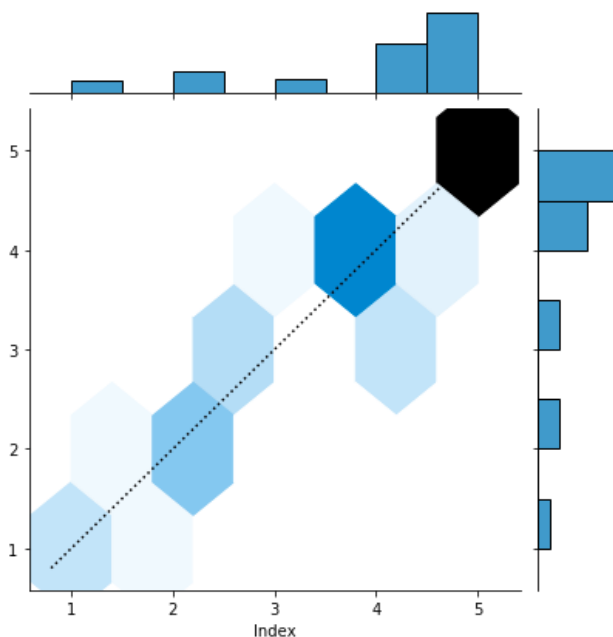

Fig 7b SVM Prediction

Fig 7c. SVM Results

The results shown above presented the performance of all the models. RandomForest, SVM, and Deep Forest. The model with the best performance in all the metrics evaluated was the Deep forest (final model selected), with 89% accuracy, 89% precision, 88% recall, and 88% F1-score and sharing the same score with the SVM model as well. The deep forest model selected even obtained values that surpass the results obtained in other similar research where obesity was analyzed with features associated with eating habits and physical condition.

V. CONCLUSION & FUTURE WORK

Being overweight and obese are regarded as epidemiological concerns due to their association with a variety of ailments, including hypertension, type 2 diabetes mellitus, osteoarthritis, stroke, some forms of cancer, and even mortality. The incidence of overweight and obesity has grown significantly on a worldwide scale, and it is anticipated that by 2030, the global prevalence of overweight would reach around 50 percent.

Obesity prevention is a difficult endeavor that needs government, business, and subject matter experts. Consequently, combating obesity and its consequences demands significant changes in the population's lifestyle, particularly in terms of physical activity and eating habits. In this context, owing to technological advancements, methods such as artificial intelligence and machine learning may be utilized to aid in the detection and prevention of overweight and obesity. Models of machine learning may be advantageous in the medical profession since experts in the field can utilize these tools as a decision-making aid.

In this study, three machine learning models, including random forest, support vector machines, and deep forest, were evaluated to develop an intelligent model for the identification of individuals with obesity or overweight, which will aid specialists in the field in their decision-making. In contrast, the data used to create the model were acquired via a survey in which information about the respondents' eating habits and physical activity was collected. The database was labeled using the BMI classification table, which used the weight and height of each instance.

GitHub link for PPT, Source code, output screenshots:

https://github.com/Vikram6630/ML-Project.git

REFERENCES

[1] World Health Organization (WHO), Obesity and overweight, 2021. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[2] H. Rosen, "Is Obesity A Disease or A Behavior Abnormality? Did the AMA Get It Right?". Missouri medicine, 111(2014): 104–108.

[3] M. Blüher, "Obesity: global epidemiology and pathogenesis". Nature Reviews Endocrinology, 15(2019): 288–298. DOI: 10.1038/s41574-019-0176-8.

[4] World Health Organization (WHO), Body mass index–BMI, 2020. URL: https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/bodymass-index-bmi

[5] T. Bhurosy, R. Jeewon, "Overweight and obesity epidemic in developing countries: a problem with diet, physical activity, or socioeconomic status?". The Scientific World Journal, (2014). doi:10.1155/2014/964236.

[6] A. Alami, A. Jafari, Z. Hosseini, "Differences in overweight/obesity prevalence by demographic characteristics and self-weight misperception status". Clinical Nutrition ESPEN, 41(2021): 249–253. doi:10.1016/j.clnesp.2020.12.005.

[7] World Health Organization (WHO), NCD risk factors: Overweight / Obesity, 2021. URL:https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-groupdetails/GHO/overweight-obesity.

[8] Kelly, W. Yang, C.-S. Chen, K. Reynolds, J. He, "Global burden of obesity in 2005 and projections to 2030". International Journal of Obesity, 32(2008): 1431–1437.

[9] R. Mehrzad, "Definition and introduction to epidemiology of obesity". Obesity, (2020): 1–6. doi:10.1016/b978-0-12-818839-2.00001-6.

[10] Y.C. Chooi, C. Ding, F. Magkos, "The epidemiology of obesity". Metabolism, 96(2018):6–10. doi: 10.1016/j.metabol.2018.09.005

[11] The Lancet Diabetes and Endocrinology, " Tackling obesity in 2020—with a great resolution comes shared responsibility". The Lancet Diabetes & Endocrinology, (2020). doi:10.1016/s2213-8587(20)30001-2.

[12] T.M. Dugan, S. Mukhopadhyay, A. Carroll, S. Downs, "Machine learning techniques for prediction of early childhood obesity". Applied Clinical Informatics, 6(2015):506–520. Doi: 10.4338/ACI-2015-03-RA-0036.

[13] K.W. DeGregory et al., "A review of machine learning in obesity". Obesity Reviews, 19(2018):668–685. doi: 10.1111/obr.12667.

[14] E. De-La-Hoz-Correa et al., "Obesity Level Estimation Software based on Decision Trees". Journal of Computer Science, 15 (2019): 67–77. doi:10.3844/jcssp.2019.67.77.

[15] M.I. Jordan, T.M., "Mitchell, Machine learning: Trends, perspectives, and prospects". Science, 349(2015):255–260. DOI: 10.1126/science.aaa8415.

[16] Hales CM, Fryar CD, Carroll MD, Freedman DS, Ogden CL. Trends in obesity and severe obesity prevalence in U.S. youth and adults by sex and age, 2007-2008 to 2015-2016. JAMA. 2018;319(16):1723–1725. https://doi.org/10.1001/jama.2018.3060.

[17] Biro FM, Wien M. Childhood obesity and adult morbidities. Am J Clin Nutr. 2010;91(5):1499S–1505S. https://doi.org/10.3945/ajcn.2010.28701B.

[18] Reilly JJ, Methven E, McDowell ZC, et al. Health consequences of obesity. Arch Dis Child. 2003;88(9):748–752. https://doi.org/10.1136/adc.88.9.748.

[19] Serdula MK, Ivery D, Coates RJ, Freedman DS, Williamson DF, Byers T. Do obese children become obese adults? A review of the literature. Prev Med. 1993;22(2):167–177. https://doi.org/10.1006/pmed.1993.1014.

[20] Hammond RA, Levine R. The economic impact of obesity in the United States. Diabetes Metab Syndr Obes. 2010;3:285–295. https://doi.org/10.2147/DMSOTT.S7384.

[21] Finkelstein EA, Trogdon JG, Cohen JW, Dietz W. Annual medical spending attributable to obesity: payer-and service-specific estimates. Health Aff (Millwood). 2009;28(5):w822–w831. https://doi.org/10.1377/hlthaff.28.5.w822.

[22] Pronk NP, Goodman MJ, O'Connor PJ, Martinson BC. Relationship between modifiable health risks and short-term health care charges. JAMA. 1999;282(23):2235–2239. https://doi.org/10.1001/jama.282.23.2235.

[23] Wolf AM. Economic outcomes of the obese patient. Obes Res. 2002;10 (suppl 1):58S–62S. https://doi.org/10.1038/oby.2002.191.

[24] Ikeda JP, Crawford PB, Woodward-Lopez G. BMI screening in schools: helpful or harmful. Health Educ Res. 2006;21(6):761–769. https://doi.org/10.1093/her/cyl144.

[25] Raczynski JM, Thompson JW, Phillips MM, Ryan KW, Cleveland HW. Arkansas Act 1220 of 2003 to reduce childhood obesity: its implementation and impact on child and adolescent body mass index.

[26] J Public Health Policy. 2009;30(1)(suppl 1):S124–S140. https://doi.org/10.1057/jphp.2008.54