Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 03 March 2024

Analyzing the Behavioral Factors Causing Obesity Using Big Data Techniques

Akhila Naga Charanya G
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil, Virudhunagar, India

Bhavana B
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education Krishnankoil, Virudhunagar, India

Madhumitha P
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil, Virudhunagar,India

Sivamurugan C
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education Krishnankoil, Virudhunagar, India

***Abstract*- Obesity has been a matter of concern for human population all over the world. For some reason, the people living in the western countries are hugely dependent on fast food to fulfill their calorie requirements. Other than that, due to great advancement in technology, they rely hugely on machines to do their manual labour which adds on to the issues related to physical fitness. This project tries to narrow down the factors that influence the level of vulnerability a person has to become obese. The data source has proved to be reliable in the past and a plethora of such analysis has been performed on data obtained from the same source. The field of big data offers a number of possibilities where different sorts of analysis are performed to make predictions which can help one to understand the patterns and trends which could easily be missed if a machine is not used. The purpose of this study is to gather data and statistics about obesity prevalence and key determinants. A foundation for comprehending the primary causes of obesity may be established by such an analysis, which will facilitate the development of a preventative and control strategy**

Keywords: **Obesity, Weight of evidence, information values,Attributes, body mass index , fisher score decision tree,continous variables,categorical variables.**

I. INTRODUCTION

A data collection that was created from a poll that collected information from more than four lakh people has been analyzed. The Centers for Disease Control and Prevention supplied data that was accessible for study in 2017. The "Behavioral Risk Factor Surveillance System" is based on information gathered between 2011 and 2015, a period of five years. The most recent data set—that is, the one related to 2015—has been analyzed for this project. Due to the large size of the data set, variables had to be chosen and outliers had to be eliminated in order to prepare it for analysis.In addition to the outliers, null values had to be removed too because the data had some missing values as well which could affect the analysis and the results adversely. Data segmentation has been done based on the variable XSTATE which tells the geographic location of the individual taking the survey. Most of the states in the US are covered in the survey and a numerical value has been given to those states to make any kind of analysis possible. The numerical values referring to the states range from 1 to 72. State wise segmentation helps in generating a predictive analysis based on the geographic location where one could easily identify the obesity trends based on various states in America.

Following the selection of variables, logistic regression is used to determine how well the predictors can forecast the target value in this example, obesity. Additionally, a decision tree has been made. A confusion matrix that indicates the prediction model's accuracy has been made in order to verify the validity of the findings. The majority of US states are included in the survey, and in order to facilitate any type of analysis, each state has been assigned a number value that corresponds to the FIPS code. The numerical values referring to the states range from 1 to 72. Hence a State wise segmentation has also been done to identify the obesity trends based on various states in America. All through the research, programming and analysis has been performing using the R language.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 03 March 2024

Figure 1: Food which causes obesity

II. RELATED WORKS

The search for the primary causes of obesity has been ongoing. Numerous studies have been conducted to determine the underlying reason ever since it was noted as a significant issue with the population. Similar to the findings presented in this paper[3], a study was conducted in 2018. That study measured a person's degree of obesity using variables like cost of meals, per capita food intake, etc. In a similar vein, a recent study indicated that fast food intake and physical activity were significant predictors of the obesity rate. It was discovered that there is a lower prevalence of obesity in locations with a high concentration of physical exercise facilities. In a similar vein, it was found that the population in locations without easy access to fast food was relatively less likely to be fat. These aspects are significant as well, but less so than the health considerations that have been taken into account for this project, such as body mass index and the existence of a condition like diabetes.

The overall goal of the investigation described in the June 2017 publication [5] was to determine the prevalence of adult males in Sri Lanka's Central Province who do not have a diagnosis of hypertension (HT). Examine the causes of undiagnosed HT in this group of people. In the Central Province of Sri Lanka, adult males have an alarmingly high prevalence of undiagnosed HT. Because of its correlation with age and BMI (weight status), it is crucial to conduct routine HT screenings and implement obesity-related treatments in order to slow the growth of this modifiable risk factor for cardiovascular disease. The relationship between fat and hypertension has long been debated. Recently, in a study of South Asian males, it has been shown that both the systolic and diastolic blood pressure is correlated to a person's age, weight, BMI, and waist circumference. Therefore, while performing feature selection for this paper, it was found that blood pressure medicines, blood cholesterol and presence of high blood pressure showed larger information value as compared to some of the other attributes in the data set. Research indicates that the incidence of diabetes, obesity, and risk factors for obesity-related health issues Reference 10 Diabetes, high blood pressure, high cholesterol, asthma, arthritis, and poor health status were all substantially correlated with being overweight or obese. Adults with a BMI of 40 or above had an odds ratio when compared to normal weight adults. Both sexes, all ages, all ethnicities, all educational levels, and all smoking levels of US citizens continue to see increases in obesity and diabetes. There are numerous important health risk factors that obesity is closely linked to.

Furthermore, the relationship between food hardship—defined as having insufficient money for food—and obesity among immigrants has been studied in a small study titled Food Hardship and Obesity in a Sample of Low-Income Immigrants [6]. This study looked at this relationship in a sample of 828 low-income, multiracial/ethnic adults in the greater Boston, Massachusetts area. Using interaction testing by place of birth, modified Poisson regression models evaluated the relationship between food hardship and obesity (BMI > 30) among people reporting hardship. A low-income population in Haiti was the subject of a research, and it was discovered that those who reported having trouble with food had a higher likelihood of being fat than those who did not. Consequently, while selecting variables for this research project, it was observed that obesity is correlated to income of anindividual as well.

III. EXISTING SYSTEM

The search for the primary causes of obesity has been ongoing. Numerous studies have been conducted to determine the underlying reason ever since it was noted as a significant issue with the population. 'An economic examination of adult obesity' was the title of a 1970s study. The analysis conducted in that study was based on differences in the cost of the food being consumed. A phone survey titled "Self-Reported Body Mass Index and Health-Related Quality of Life" was used to collect data from US citizens. Only adults were included in the data collection. The physical component of BMI was taken into account in this study, however the survey's sample size was limited .Only adults were included in the survey which is a limitation as many children also face the problem of obesity nowadays.

Figure 2: consequences, reasons and preventions

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 03 March 2024

IV. MEHODOLOGY

The analytic project can be diving into six phases as follows:

A. Data collection

The project's first phase is gathering data so that analysis may be performed. The Centers for Disease Control and Prevention (CDC) submitted the data collection used in this work to the open platform Kaggle in 2017. Data has been produced with the use of survey data collected from more than 400,000 US citizens. The survey was conducted between 2011 and 2015, spanning a five-year period.

Data pre-processing Phase two of the project aims at making the data set ready for analysis by removing the null values and making sure the values of all the attributes are numerical ones.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 03 March 2024

Table 1:Literature survey of research papers

Title of researchpaper	Objective	Methodology	Advantages	Disadvantages
"Associations betweenfast food and physical activity environments and adiposity in mid- life: cross-sectional, observational evidencefrom UK Biobank" (Mason et al.,2018)	To examine the association between exposure to fast food outlets, physical activity opportunities,and an individual's adiposity levels.	To examine the association between exposure to fast food outlets, physical activity opportunities,and an individual's adiposity levels.	- Large sample size - Comprehensive data collection - Ability to identify potential causal relationships -	- Cross-sectional design limits inferring causality -Potential for confoundingfactors
"Genetic and environmental determinants of BMI and obesity in youth: areview" (Lobdell et al.,2017)	To review the current understanding of the genetic and environmental factors that contribute to bodymass index (BMI) andobesity in youth.	Comprehensive literature review of studies investigating the genetic and environmental influences on BMI andobesity in children andadolescents.	- Provides a comprehensive overview of the latest research - Highlights the complex interplay between genetics and environment	- Focused on childhood obesity, may not directlyapply to adult obesity - Limited to observationalstudies
"The role of early life environment in the development of obesity" (Whitaker etal., 2016)	To explore how early life experiences, such as maternal obesity, breastfeeding duration,and exposure to processed foods and sugary drinks, can influence the risk of developing obesity later in life.	Review of epidemiological studies and clinical trials examining the association betweenearly life factors andobesity risk.	- Highlights the critical role of early life factors in obesity prevention - Provides insights into potential intervention strategies - Identifies modifiable risk factors in early life	- May not be able to fully disentangle the complex interplay of early life experiences and obesity risk
"The influence of sleepon appetite, metabolism, and obesity" (Chaput et al.,2016)	To investigate the relationship between sleep duration and quality and its impacton appetite, metabolism, and the risk of obesity.	Review of experimental studies and observational studies examining theeffects of sleep on appetite, energy expenditure, and obesity risk.	- Provides evidenceof the importance of sleep for obesity prevention and weight management - Identifies potential mechanisms linking sleep and obesity	- Focuses on sleep as arisk factor, may not provide comprehensivestrategies for sleep improvement
"The role of stress inobesity: an update" (Epel et al., 2009)	To summarize the current research on the role of stress in obesity, with a focus on the mechanisms by which stress can promote fat storage	Review of experimental studies and observational studies examining theeffects of stress on hormones, metabolism, and	- Provides insights into the physiological mechanisms linkingstress and obesity - Identifies potentialtargets for	- Focuses on stress as arisk factor, may not provide comprehensivestrategies for stress management

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 03 March 2024

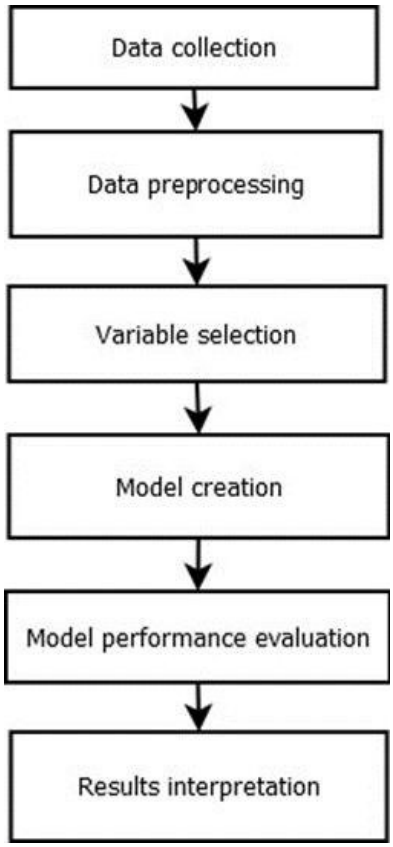

Figure 3: Phases of the project

B. Variable selection

The original data set is huge as it contains a lot of attributes and covers the collected data from all the five years the survey has been carried out for. Many of the attributes like telephone number, interview month, address, and so on are not required for the kind of analysis needed to be done for this project.

Therefore, in the fourth phase of the project, variable selection is done using weight of evidence, information value and fisher score. Our data is a mixture of continuous and categorical variables. The selected target variable for prediction is categorical in nature which is OBESITY. Information value is calculated for categorical predictor variable against the categorical target variable. Information value is interpreted using the categories given in the following table:

Table2: Predictive power of variables based on IV

Information Value	Predictive power
<0.02	Useless for prediction
0.02 – 0.1	Weak
0.1 – 0.3	Medium
0.3 – 0.5	Strong
>0.5	Too good to be true

Fisher Score is calculated for continuous predictor variable against the categorical target variable.The variables with suitable values which can be better observed from the * representation of significance, for instance the variable BMI

with "***" significance has been selected as it represents a higher predictiveness and variable with "*" significance has been ignored as it represents weak predictiveness

After analyzing the values we have filtered the variables in accordance to their prediction power. For continuous variables glm() function was used over the dataset to select the variables.

Weight of evidence is a metric which is used to prove or disprove a certain hypothesis. It is computed as follows:

```
WEIGHT2      1.765e-04  6.951e-05   2.539 0.011121 *
HEIGHT3     -1.347e-04  4.380e-05  -3.075 0.002108 **
BMI          6.000e-04  4.722e-06 127.069  < 2e-16 ***
FRUIT1       3.030e-05  8.989e-06   3.371 0.000749 ***
```

Figure 4: Fisher Scores of selected variables

$$WoE = \ln(DistributionGood/DistributionBad) \quad \ldots (i)$$

A positive WoE value means that in the group DistributionGood is greater than DistributionBad and negative WoE means that DistributionBad is greater than the DistributionGood. Using the WoE thus calculated, Information Value is also calculated as explained below:

$$IV = \sum (DistributionGood_i - DistributionBad) \times WoE \ldots (ii)$$

Seventeen characteristics were chosen to be included in the building of the model, which will be further explored in the findings, based on the Information Values and the Fisher Score produced for various variables of the year 2015.

D .Model creation

Phase four of the project consists of creating prediction models. The models generated after the feature selections done in phase three include logistic regression and decision tree.

E. Model performance evaluation

The models are evaluated in phase five by calculating the accuracy derived from the confusion matrix.

F. Results interpretation

In the end, the results of the predictions made by the models are interpreted which could help one to draw conclusions about the behavioral factors which influence obesity the most.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 03 March 2024

Table 3: Attributes selected based on the information valueS

S. No.	Variable	IV
1	DIABETE3	0.3242424
2	GENHLTH	0.1239037
3	EMPLOY1	0.03271449
4	HAVARTH3	0.03225654
5	INCOME2	0.03142882
6	EXRACT11	0.02353044
7	BPMEDS	0.01680554
8	HLTHPLN1	0.01594627
9	BLOODCHO	0.01552571
10	CHCKIDNY	0.01476271
11	BPHIGH	0.01388662
12	HRTATK	0.01338881
13	X_STATE	1.67307E-05

V. RESULTS

After selecting suitable features using Fisher score and information value, prediction models have been created. To create the logistic regression model, data has been divided in the ration 3:2 for training and for testing respectively. After the model is created, its accuracy has been derived from the confusion matrix.

The accuracy of the model comes out to be 87.5028 percent. Despite of having no predictive power X_STATE is still considered for further reference for segmentation. As shown in figure3, the decision tree model classifies a person to be obese or not based on factors like height and weight.

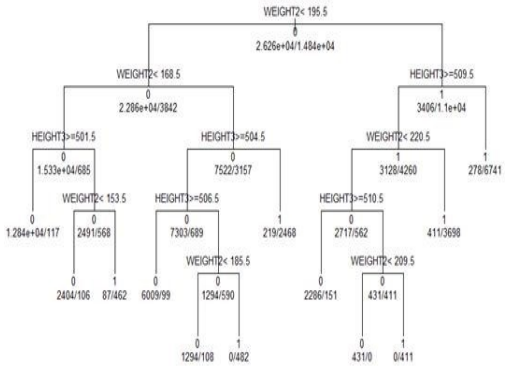

Figure 5: Decision tree for performing classification

If a person is not classified as obese, the leaf node of the tree produces an output of 0, and if they are, it produces an output of 1. Every time a division occurs, the model receives the values of a variable and, using a predetermined threshold, divides the data The model receives inputs from a variable, divides them based on a predetermined threshold, and then either produces another node or makes the ultimate decision. Let's take an example where the person is 180 pounds and 165 centimeters tall.

According to the root leaf, it is less than 195.5 so at the next point of bifurcation the weight is compared to the value 165.5. now 180>165.5, so the next value which serves as a point of bifurcation is the person's height. now the height of the individual is 5.4 feet which categorizes him/her as obese.

CM	Predicted 0	Predicted 1	Total
Actual 0	24728	1180	25908
Actual 1	3846	10463	14309
Total	28574	11643	40217
Accuracy:	87.5028		

Figure 6: Confusion Matrix

The confusion matrix, which is seen in Figure 6, is produced when the model is developed and shows the number of true positives, true negatives, false positives, and false negatives predicted by the model.

Figure 7 uses a color gradient from white to red to visually represent the prevalence of obesity in each state, with bright red representing the most prominence and white representing the lowest. The states that are shown in gray are those that were left out of the poll. Initially, a data frame is built that connects the variables STATE and OBESITY. Subsequently, an additional data frame is constructed to contain the total frequency counts of obesity for every state value. Plotting the map and determining the gradient—which increases as OBESITY frequency counts rise—uses this data frame containing the frequency counts of OBESITY and the FIPS codes included inside the STATE variable as references.

the state where the greatest number of cases of obesity are reported. According to figure 7, the state of Florida has the greatest number of cases reported for obesity.

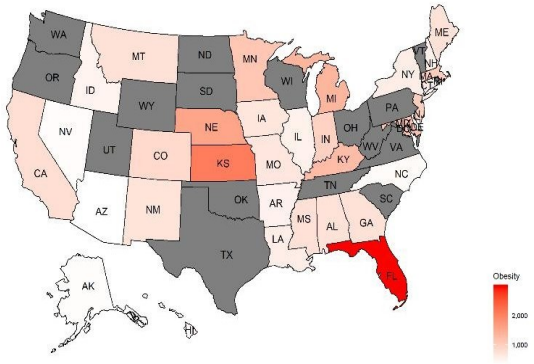

Figure 7: State wise prominence of obesity

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 03 March 2024

VI.DISCUSSIONS

The model handles the factors that may affect the probability of one becoming obese and achieves it, but the data selected mainly only focuses on behavioral factors and ignores the physical and genetic factors that might also contribute towards a person's possibility of getting obese.

The state-wise analytic approach is highly generalized. The survey data does not contain regional factors nor takes the persons origin into the consideration and only consists of behavioral factors. Hence state-wise analytics of obesity is legitimate but lacks dependability as it cannot be considered as a deciding factor.

VII . CONCLUSION

The aim of the project was to find the relationship of various factors which trigger obesity in a person. Initially a huge number of attributes were available and, in the end, after feature selection, only 17 attributes were considered which somehow influence the susceptibility of a person to becoming obese. Various classification models were created, and their accuracies were noted. As a result, it was noted that factors like a person's general health, BMI, weight, height, blood pressure, continuous intake of blood pressure medication and if he is diabetic or not, influence his/her course to becoming obese. Out of all the attributes known, obesity can at best be associated with these said factors

VIII.FUTURE ENHANCEMENTS

This project has a very good scope in future. Finding the factors which trigger obesity in an individual can be easily done, if the survey is performed successfully in every area. This project can be helpful to the Healthcare centers. Thus, helping individuals in prevention and control plan.

IX . REFERENCES

[1] S. Chou, M. Grossman, H. Saffer, "An economic analysis of adult obesity: results from the Behavioral Risk Factor Surveillance System", Elsevier, Vol 23, May 2004.

[2] E.S. Ford, D.G. Moriarty, M. Zack, A.H. Mokdad, D.P. Chapman, "Self-Reported Body Mass Index and Health-Related Quality of Life: Findings from the Behavioral Risk Factor Surveillance System", Wiley online library, September 2012.

[3] K.E. Mason, N. Pearce, S. Cummins, "Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank", NCBI, January 2018.

[4] G. Manogaran, D. Lopez, "Health data analytics using scalable logistic regression with stochastic gradient descent", Inderscience, vol 10,2017..

[5] N.W.I.A. Jayawardana, W.A.T.A. Jayalath, W.M.T. Madhujith, U. Ralapanawa, R.S. Jayasekera, S.A.S.B. Alagiyawanna, A.M.K.R. Bandara, N.S. Kalupahana, "Aging and obesity are associated with undiagnosed hypertension in a cohort of males in the Central Province of Sri Lanka: a cross-sectional descriptive study", BMC cardiovascular disorders, June 2017..

[6] C.E. Caspi, R.D. Seely, A. Gary, C.A. Roberto, A.M. Stoddard, "Food Hardship and obesity in a sample of low- income immigrants", Journal for Immigrant and Minority Health, Vol 19, February 2016.

[7] Alexis Ruffault, Sebastien Czernichow, Martin S. Hagger, Margot Ferrand, Nelly Erichot, Claire Carette, Emilie Boujut, Cecile Flahault, "The effects of mindfulness training on weight-loss and health-related behaviours in adults with overweight and obesity: A systematic review andmeta analysis".

[8] Luke N Allen, Jessica Pullar, Kremlin Khamarj Wickramasinghe, Julianne Williams, Nia Roberts, Bente Mikkelsen, Cherian Varghese, Nick Townsend, "Evaluation of research on interventions aligned to WHO 'Best Buys' for NCDs in low-income and lower-middle-income countries: a systematic review from 1990 to 2015".

[9] Farshad Firouzi, Amir M. Rahmani, K. Mankodiya,M. Badaroglu, G.V. Merrett, P. Wong, Bahar Farahani, "Internet-of Things and big data for smarter healthcare: From device to architecture, applications and analytics".2018.

[10] Ali H. Mokdad, Earl S. Ford, Barbara A. Bowman, "Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors", division of adult community health,2001

[11] JT Stadler, G Marsche, "Obesity related changes in high density lipoprotein metabolism and function", International journal of molecular sciences,2020

[12] E Shams, V kamalumpundi J Peterson, "Highligshts of mechanism and treatment of obesity related hypertension" Journal of human hypertension ,2022

[13] M Wlodarczyk, G Nowika ," Obesity, DNA damage and development of obesity related diseases International journal of molecular sciences,2019