Influential Factors Determining Obesity and Fitness

1st Gaurav Gupta
dept. of Computing
National College of Ireland
Dublin, Ireland
x22212311@student.ncirl.ie

2nd Vipin Sharma
dept. of Computing
National College of Ireland
Dublin, Ireland
x22207406@student.ncirl.ie

3rd Himani Sharma
dept. of Computing
National College of Ireland
Dublin, Ireland
x22224815@student.ncirl.ie

***Abstract*—Meeting the multifactorial problem of obesity will be complicated and therefore, it will need multilevel solutions that will address the social, economic, and environmental factors. Machine learning as one of the solutions provides a basis for predetermining obesity levels and processing customized interventions. After a thorough review of papers in this domain, researchers applied various machine learning models like logistic regression, decision trees, and random forests, predictive models to assess obesity risk. CRISP-DM methodology is the basis of the project which includes matchmaking and aligning with business targets. In qualitative analyses, the diagnosed cases are reviewed by age, level of physical activity, and history, which give us specific insight into the targeted intervention. For this study, we use the Random forest classifier to predict the Obesity level based on important features such as weight, height, Family history with overweight, and smoking habits. Fundamentally, the addition of machine learning to obesity research and the application of intervention strategies can certainly assist in combating the number one public health problem that is quite prevalent.**

***Index Terms*—K-Nearest Neighbor, Random Forest, LR, F1-Score, Obesity**

I. Introduction

Meeting the obesity challenge in the global society is formidable because it is the multi-layered interaction of genetic redispersion, social, economic, and environmental factors. Through more and more people being obese, the risk for long term diseases such as heart disease, diabetes, and some cancers is also increasing. Obesity is on a prevailing wave again. However, the most concerned matter is the growing number of overweight kids caused by wrong foods, lack of activity, and a surrounding signalling for sweet snacks. Reducing obesity calls for an inclusive strategy including policy measures that are aimed to create a healthy environment, implementing fair systems that provide healthy food access and the provision of opportunities for physical activities. Personal responsibility is indisputably great, even for new solutions in the battle against this pressing issue but acknowledging and trying to remove the structural causes is equally, if not more, important. The model can also be applied for the personalized training programs and the fitness industry. Such findings can be applied to most informed strategies including, but not limited to, nutrition guides and product recommendations based on meal frequency, calorie intake, and consumption of vegetables among others. Besides that, the model can allow trainers to keep an eye on their customers' activity, including the usage of electronic gadgets, the amount of water they drink and provide feedback and goal setting. Going forward, the model might develop into something that will have more functions like heart rate monitoring, activity levels as well as sleep patterns and other important parameters. Apart from fitness, this predictive model could be implemented in healthcare, through BMI calcs calculation, prediction of diseases like diabetes/cardiovascular issues, and addressing the mental health problems associated with obesity.

II. Hypothesis

Null hypothesis - is there any correlation between obesity and factors such as family history, smoking, food habits, physical activity

Alternate hypothesis - there is no correlation between any of such factors and obesity

III. Literature Review

According to paper [1], Feature selection has not been done. Evaluate the model with mean squared error, mean absolute error, and R-squared. Correlation to find the relationship between variables ANOVA testing to check the hypothesis between age and BMI. They have used the SVM, Naive Bayes, Decision Trees, Random Forest, and Logistic Regression. Refering this Paper [2] states that author used a LSTM(a type of recurrent neural network) used to capture long and short-term dependency. However, the dataset is very restricted to the pediatric population and also increases the model complexity but accuracy remains constant. The paper [3] also took the dataset related to children and used machine learning models such as RandomTree, RandomForest, and Naïve Bayes. Still, some of the methodologies are missing which can give better accuracy and missing on attribute selection. We extracted from Paper [4] that research gives solutions using physical activity data monitoring. Results demonstrated the direct effects of physical activity on the prediction of the weight status and specific activities playing an important role. Demographic features such as gender and age did contribute to the results as well. The results emphasize the matter of appropriately designed programs for the prevention of obesity rates differentiating the population. According to paper [5], ML techniques provide a powerful tool for obesity predictions via the use of algorithms such as regression, neural networks, decision trees, and ensemble methods to determine body fat, identify predictors, and forecast obesity prevalence. These methods

help in the prevention and countering of prevalent health problems.

According to Paper [6], This review article explains the complicated relationship between obesity and a wide range of diseases, which have a very negative effect on the health of the society. It, with great attention, examines the medical ramifications of obesity, from such diseases as the cardiovascular and respiratory disorders, to neurodegenerative conditions and the autoimmune disorders. Besides, the review provides an insight into application of machine learning (ML) techniques in obesity prediction, and the identification of powerful approaches like artificial neural networks. The systematic approach to research questions, as well as the clinical implications discussed, signals for the importance of ML in obesity detection and treatment. Although database restrictions restrict, the review provides useful information and paves the way for the future and studies that aim to understand and combat the world's epidemic of obesity and its constituent health hazards. From Paper [7], The research employed a wide analysis method via the "BMI" dataset that consists of 500 records and feature like gender, height, weight, and index. Our study applied multiclass classification and aimed to group the people according to their body composition, which had a high accuracy of 95 percent with the spherical linear kernel SVM model. Analysis of multi classes data set, binary classification was done to predict whether an individual is obese or not. Classifiers using SVM without calibration and above-mentioned Decision Tree classifier equally performed well in terms of binary classification, but the last one especially after it was calibrated following the "isotonic" method did really well. Later on, examining the "insurance" dataset, which compiled 1338 records, unveiled similar links between smoking and the insurance payments. As to SVM and Decision Trees, these two classifier algorithms were proved to work best, with a good level of performance in the case of only a plane amount of information. The investigation pinpointed some risk factors directly linked to obesity and overweight, which include BMI, age, or tobacco use, dietary habits, and socioeconomic factors. Finally, the article addressed future study objectives placing an emphasis on creating an intelligent eCoach system that targets obesity and helps in achieving personal wellness goals. This process is focused on providing the early forecasts and the recommendations which are individual-oriented and the basis of gathered data. According to Paper [8], The review discusses a detailed comparison of the performed statistical and machine learning analyses to forecast the overweight and obesity of the child and adolescents. While the statistical methods like logistic regression is useful for their simplicity and interpretability, the ML algorithms such as ANNs(artificial neural network) and deep learning are the best techniques to handle the complex, high-dimensional data that are uncovering the nonlinear relationships. ML models perform better than statistical models in the prediction of outcomes. It is especially valid in the application of multiple covariables with new-generation risk factors along with old risk factors. In spite of the fact that they are known as black boxes, it contain interpretability enhance techniques, for example, variable importance, attention mechanisms and reinforcement learning. The review, therefore, brings out the ML techniques as the emergent paradigm in the obesity research domain with the capability to revolutionize obesity prevention efforts and is therefore of value for the studying community as well as practitioners. Referring to paper [9], This section is devoted to the study whose goal is obesity susceptibility prediction by involving genetic data of DTCGT participants from the Personal Genome Project (PGP) to analyze. A set of seven machine learning algorithms is investigated for their accuracy employing sensitivity and specificity as measures of performance. Firstly, some genetic variants which are associated with obesity are found and reduced using random forest, as they are the biggest data set. Thirteen SNPs that relate to obesity and other characteristic diseases are later picked for the prediction. Models are developed and tested using R language, in which support vector machines showed great capability, becoming the most efficient classifier with high AUC (area under the curve) values. The outcome of the study on machine learning pointing to forecasting of complex diseases justifies the role of personalized patient care. There should be detailed research on model specific feature selection techniques for the present purposes.

IV. Methodology

The framework used in this project is CRISP-DM as it provides better business understanding, objectives, and goals. It is also helpful in assessing the current situation and developing project plans.

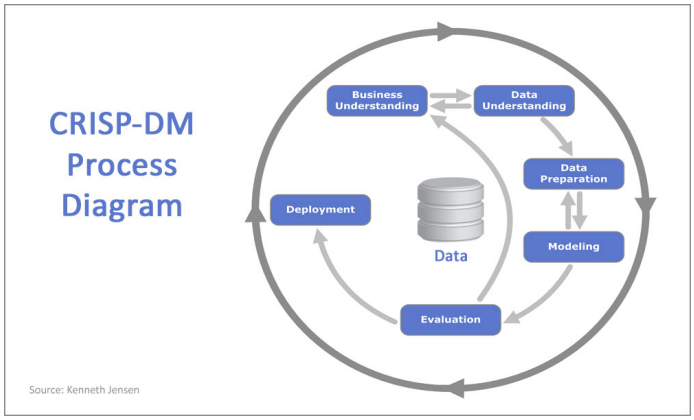

Fig. 1. CRISP-DM Methodology)
https://stellarconsulting.co.nz/wp-content/uploads/2017/08/CRISP-DM_Process_1000x600.jpg

A. Business Understanding

The healthcare and fitness industry needs a large amount of data to recommend the right products to them. This data can also be useful for a fitness organization for one-to-one training as well as suggesting a diet chart and plan through their mobile application or web portals. This can also be useful by hospitals for potential risks to the population.

B. Data Understanding

The dataset is from an open source platform which is UCI machine learning repository. It is a widely used platform in the machine learning and Data science community. The dataset consists the information about obesity levels of different age groups people. The data consists of 2112 unique customers with 17 independent variables such as Gender, Age, and Family History of overweight and only one dependent variable which we have to predict.

C. Data Preprocessing

The Preprocessing steps are the heart of the model we have to perform many things on the dataset so that in the end model will give good results. Once we understand the data and then loading the dataset as a CSV file, the first step is to check missing values and handle it with the mode of that value. Checking the outlier with z-score test as data is discrete so there are no outliers such that. Transforming the data by normalizing and standardizing. Transforming the data by min-max scaler. Converting the categorical data into numerical using label encoding so that it can be better for machine learning algorithms. Now our data is ready to apply machine learning algorithms.

D. Data Modelling

- **Random Forest Classifier**: There are many traditional Machine learning Models deployed in this domain as we see in the literature review section for this project we are using the Random Forest Classifier. [12]Random forest classifier is a very effective tree learning algorithm in machine learning. During the training phase,As shown in Fig.5, it will create many Decision Trees and each tree is built by using a random subset of features in each segmentation. The randomness of all separate trees will reduces the overfitting, increase the accuracy, and improve the overall prediction performance of the model. Random forest is mostly used for the classification and regression. It can also handle the missing value very well without performing the imputation and the best part of this model is it can maintain the accuracy of the model even when the large amount data is missing from the dataset.

V. Results/ Finding

A correlation matrix is a relationship between several variables in a dataset. Each cell of the square matrix that represents the correlation is a correlation coefficient, ranging from -1 to 1, linear relationship between two different variables. The coefficient equals to 1 means a perfect positive correlation where two variables change the same way. -1 designated is perfect negative correlation where one variable goes down when the other variable goes up. We can conclude that a coefficient around 0 implies a slight to no linear relationship between variables under evaluation. Therefore, while correlation and covariance matrices are definitely of great help in the process of interpretability of data, they should still be used carefully

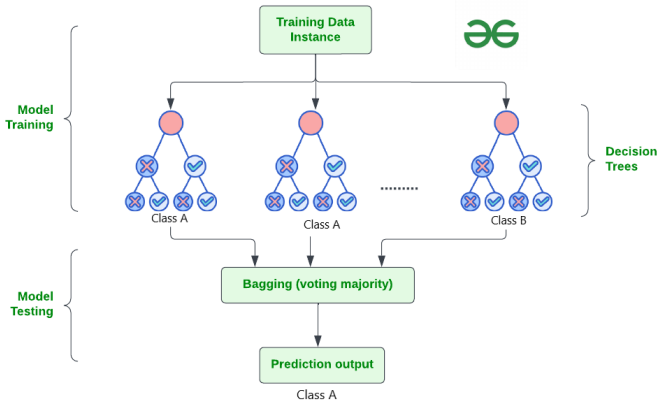

Fig. 2. [12]Random Forest Algorithm Diagram

and always in parallel with other analytical techniques for the sake of abstracting useful and conclusive information. Here, we can see none of the factors contributing that much but all the factors have significance contribution. Have created a correlation matrix which has an age as a negative correlation. The Fig 3. shows better visualization matrix

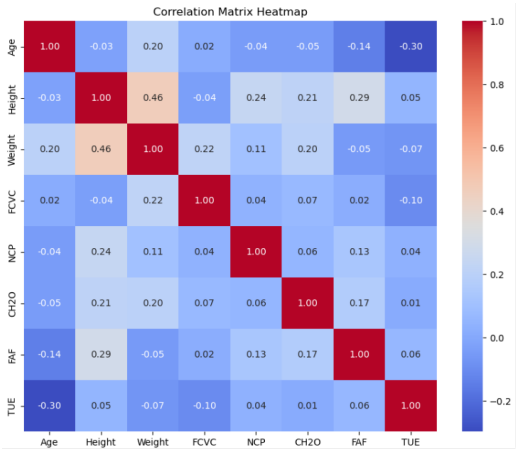

Fig. 3. Heatmap

In Figure 4 explains that According to experts' opinion, genetic defects are very prominent in childhood and adolescent obesity. On the other hand, emotional disorders such as obesity are largely contested as one of the contributing factors of obesity by a large proportion of the respondents. It is a noteworthy percentage of the respondents who held that the major source of obesity was one's work.

Figure 5 also explains the predominant number of respondents soundly denies the idea that a lack of physical activity is the greatest cause of overweight. While that's the case, smoking cigarettes usually is not considered one of the main causes that lead to obesity.While doing specific types of work indicates the major element among the causes of being overweight.

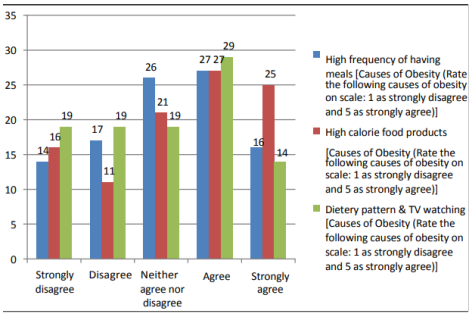

Fig. 4. [10]Causes of Obesity

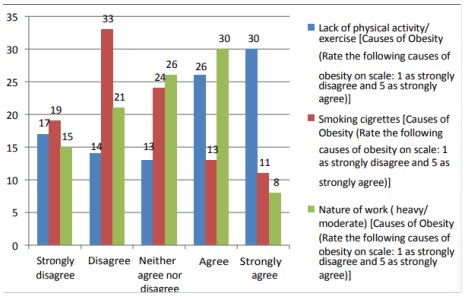

Fig. 5. [11]Cause of Obesity

VI. Interpretation of the Results

	precision	recall	f1-score	support
Insufficient_Weight	0.98	0.96	0.97	51
Normal_Weight	0.84	0.98	0.91	54
Obesity_Type_I	0.91	0.94	0.92	63
Obesity_Type_II	0.98	0.89	0.93	56
Obesity_Type_III	1.00	0.99	0.99	74
Overweight_Level_I	0.97	0.87	0.92	71
Overweight_Level_II	0.88	0.93	0.90	54
accuracy			0.94	423
macro avg	0.94	0.94	0.94	423
weighted avg	0.94	0.94	0.94	423

Fig. 6. Model-1 Random Forest Classifier

After Reviewing all the papers we have decided to use the Random forest classifier on this dataset because of its Versatility and Robustness. When applied this model we got good insights and accuracy. The precision is between 0.84 to 1.00 which measures the accuracy of positive prediction and the value of precision shows that the model correctly identifies instances of each class. The Recall value ranging from 0.87 to 0.99, it measures the proportion of correct predicted value among all the classes like Insufficient Weight class has the highest recall 0.96 which shows the 96% of actual values were correctly identified. The next evaluation metric is the F1-score which ranges between 0.90 to 0.99 for all the categories it also show good results. Comes at the end the overall accuracy of the model in 94% which indicating that the model performs very good and It can accurately classify people into their respective Obesity classes or different levels.

VII. Conclusion

To conclude, after applying the Random Forest classification model to our Obesity dataset has given the valuable insights and given the good results on the targeting column. Among all the models which we studied in the literature review section the Random Forest classifier performed very well with very high accuracy 94%. These findings give the value of ensemble-based techniques such as Random Forest and the adaptability of support vector machines in classification applications. Also by looking at the Correlation heat map it clearly shows that there is a correlation between obesity and independent variables such as family history, smoking, and food habits. So we can say that the null hypothesis is accepted and the alternate hypothesis is rejected which shows there is no correlation between the variables.

References

[1] Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2020). Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview. *Sensors, 20*, 2734. https://doi.org/10.3390/s20092734

[2] Cheng, E.R., Steinhardt, R., & Ben Miled, Z. (2022). Predicting Childhood Obesity Using Machine Learning: Practical Considerations. *Biomedinformatics, 2*, 184–203. https://doi.org/10.3390/biomedinformatics2010012

[3] Cheng, X., Lin, S.-y., Liu, J., Liu, S., Zhang, J., Nie, P., Fuemmeler, B.F., Wang, Y., & Xue, H. (2021). Does Physical Activity Predict Obesity—A Machine Learning and Statistical Method-Based Analysis. *International Journal of Environmental Research and Public Health, 18*(8), 3966. https://doi.org/10.3390/ijerph18083966

[4] Dugan, T. M., Mukhopadhyay, S., Carroll, A., & Downs, S. (2015). Machine Learning Techniques for Prediction of Early Childhood Obesity. *Applied Clinical Informatics, 6*, 506–520. http://dx.doi.org/10.4338/ACI-2015-03-RA-0036

[5] Safaei, M., Sundararajan, E. A., Shapi'i, A., Driss, M., & Boulila, W. (2021). A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Computational and Biomedical Research, 104754.* https://doi.org/10.1016/j.compbiomed.2021.104754

[6] M. Safaei, E. A. Sundararajan, A. Shapi'i, M. Driss, and W. Boulila, "A systematic literature review on obesity: Understanding the causes and consequences of obesity and reviewing various machine learning approaches used to predict obesity," Computers in Biology and Medicine, vol. 137, p. 104754, 2021. Available: https://doi.org/10.1016/j.compbiomed.2021.104754

[7] Yagin, F. H., Gülü, M., Gormez, Y., Castañeda-Babarro, A., Colak, C., Greco, G., Fischetti, F., and Cataldi, S. (2023). Estimation of Obesity Levels with a Trained Neural Network Approach Optimized by the Bayesian Technique. Applied Sciences, 13(6), 3875. https://doi.org/10.3390/app13063875

[8] Colmenarejo, G. (2020). Machine Learning Models to Predict Childhood and Adolescent Obesity: A Review. Nutrients, 12(8), 2466. doi:10.3390/nu12082466

[9] Montañez, C. A. C., Fergus, P., Hussain, A., Al-Jumeily, D., Abdulaimma, B., and Hind, J. (2017). Machine Learning Approaches for the Prediction of Obesity using Publicly Available Genetic Profiles. In 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp. 2743). IEEE. DOI: 10.1109/ATSIP.2017.8075569

[10] ”Causes of Obesity,” ResearchGate. Available: https://www.researchgate.net/figure/Causes-of-Obesity_fig1_355476155
[11] ”Causes of Obesity,” ResearchGate. Available: https://www.researchgate.net/figure/Causes-of-Obesity_fig1_355476155
[12] GeeksforGeeks, “Random Forest Algorithm in Machine Learning,” GeeksforGeeks, Feb. 22, 2024. [Online]. Available:https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/