Predicting Obesity Classifications from Lifestyle factors through a Machine Learning Approach.

Obesity is a growing epidemic (Safaei *et al.*, 2021)**.** Obesity is linked to many health implications. Many commonly known complications from obesity including cardiovascular diseases, cancers (Prospective Studies Collaboration *et al.*, 2009) and Type 2 Diabetes (Eckel *et al.*, 2011). Obesity has also been linked to many other health conditions affecting almost very system in the human body (Kinlen, Cody and O'Shea, 2018; Bischoff *et al.*, 2017).

Many countries are affected by the cost of obesity on their health services (Tremmel *et al.*, 2017). Increased obesity prevalence has been seen to significantly increase the cost of healthcare (Andreyeva, Sturm and Ringel, 2004). Predicting the proportion of the population with obesity, or at risk of developing obesity can provide decision makers with valued information to aid in financial and resource allocations. Many factors have been seen to be linked to the progression of obesity, such as genetics, socioeconomic factors, family history and lifestyle factors such as eating habits and activity levels (Williams *et al.*, 2015). Accessing this information at a population and individual level can determine the likelihood of obesity development.

There are previous studies that build machine learning approaches involved in obesity prediction (Uçar *et al.*, 2021). However it is stated that there are few studies that develop models beyond simple models and further development into creating obesity prediction models is needed (Safaei *et al.*, 2021). Building models which predict obesity from lifestyle factors could have uses in healthcare settings. These would allow healthcare providers to predict individuals at potential risk of obesity or individuals at risk of developing further health consequences from obesity. This can be important in preventative healthcare, using treatment to stop the development of a disease. This would aid in curbing the cost of obesity on healthcare systems.

Data

The data used (Palechor and Manotas, 2019) for this classification problem was collected though a web survey from individuals in Mexico, Peru and Colombia. It contains data on obesity level and the lifestyle habits. The dataset used is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The categories for the obesity levels were calculated through body mass index (BMI), see equation 1, and the standards set out by the World Health Organization (WHO) and Mexican Normativity (Palechor and Manotas, 2019). The original researchers used a SMOTE method (Chawla *et al.*, 2002) to balance the classes from the data collected. This resulted in 77% of the data being synthetically generated. This was performed to allow the subsequent development of predictive models from this data (Palechor and Manotas, 2019). As the original researchers performed preprocessing techniques on the data, missing values would have been removed prior to this, the data was double checked for missing values before the model building.

Exploratory Data Analysis

The height and weight values show a moderate correlation of 0.46, with a p value less than 0.05, showing that this is significant. Figure 1 shows the relationship between Height and Weight.

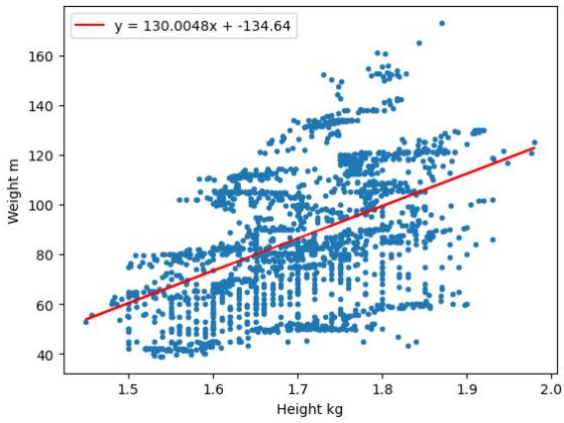

Figure 1. Scatter plot of Height against Weight. The regression line shows the linear relationship between the two variables.

The height and weight variables were combined to create the variable BMI (body mass index). This is defined by the equation 1.

$$BMI = weight / height^2$$

The boxplot in Figure 2 shows there are no outliers in the calculated BMI variable.

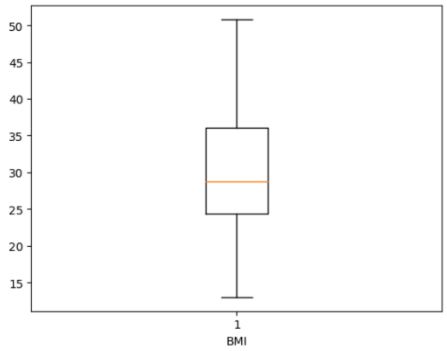

Figure 2. Boxplot showing BMI distribution.

Exploratory data analysis was performed to analyse the affect of the different variables on the classification of obesity level. Figure 3 shows the mean levels of physical activity for each target group. The bar chart shows that there is a decreasing trend in the levels of physical activity as the level of obesity increases.

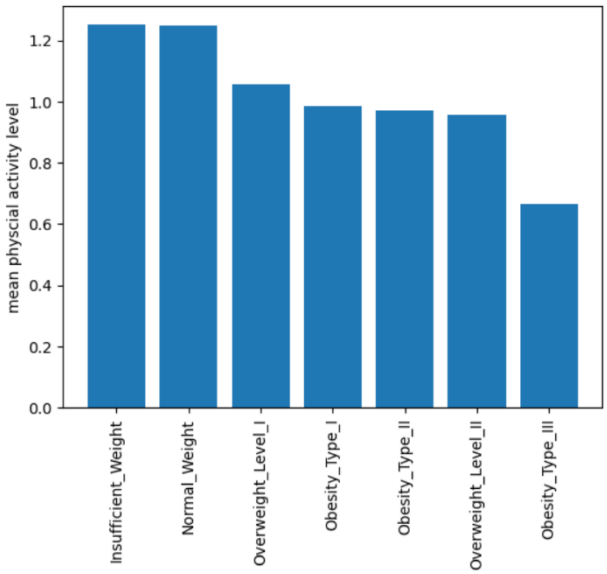

Figure 3. Bar chart showing the mean physical activity for each obesity level.

family_history_with_overweight	no	yes
Target		
Insufficient_Weight	146	126
Normal_Weight	132	155
Obesity_Type_I	7	344
Obesity_Type_II	1	296
Obesity_Type_III	0	324
Overweight_Level_I	81	209
Overweight_Level_II	18	272

Table 1. A contingency table showing the frequency of the observations in the dataset with a family history of obesity for each obesity level.

Chi-squared statistic	**P-value:**
621.98	4.2280167944702657e-131

Table 2. The results of the chi-squared test for the contingency table in Table 1.

Table 1 shows the relationship between the target variable (obesity level) and the frequency of observations with a family history of obesity. As the obesity level increases, table 1 shows that there

is an increase in the difference between people reporting a family history of obesity and those without a family history of obesity. The result of the chi-squared test in Table 2 shows that there is a significant association between these variables.

Data preparation for Machine Learning Models

Prior to performing the machine learning algorithms, the BMI column was dropped. This was due to the categories for the obesity levels being determined by the BMI score. This means the BMI was the same as the target variable in a continuous format. The categorical features of the data were encoded to allow the data to work with machine learning models. Binary variables were encoded to make the values 0 and 1, ordinal variables were encoded with values to allow the order of the categories to be maintained. The single nominal variable, MTRANS (Transportation used) was encoded using one-hot-encoding to ensure no order was introduced into the categories of this variable. The data was split into training, test and validation sets with a split of 70%, 15%, 15% respectively. Normalisation was used to ensure all features were on the same scale. This prevented features with larger values from having an inflated effect on the models compared with features with smaller values.

Models

Supervised learning algorithms were chosen for this problem as the goal is to classify the observations into one of seven categories. Five different algorithms were trained and tested for this classification problem. A grid search approach was used during hyperparameter tuning for the models, during which a five-fold cross validation was performed to analyse the performance of the model with each hyperparameter combination. The training set was used to train the models and the testing set was used to test and tune the hyperparameters. The validation set was not touched throughout this process to prevent data leakage. This set was saved to evaluate the performance of the best model using unseen data.

Support Vector Classifier (SVC) is a model that finds a hyperplane to separate the classes in the data. In this problem, the algorithm is adjusted for multi-class classification by setting it to compare one-vs-rest. This model can be computationally expensive with many hyperparameters to tune. The hyperparameters tuned in the SVC classifier model were kernel, C and gamma. The hyperparameters chosen from the grid search were rbf for kernel, 100 for C and 1 for gamma.

K-nearest neighbours works by classifying a new observation based on the classifications of a particular number of neighbours. These models are easy to interpret, but can be computationally expensive to tune the hyperparameters, primarily calculating the k (number of neighbours) for large datasets. The hyperparameters chosen from the grid search were Manhattan for the distance metric and 1 for k.

Random Forest algorithm uses bagging to create decision trees and aggregates them. The hyperparameters chosen from the grid search were entropy, for the criterion, and 200 for the number of estimators.

Adaptive Boosting is an ensemble method, it can be used with many base algorithms, decision trees have been chosen for this model. It works by performing multiple of the base algorithms and assigning them a weighting based on the number of errors made, these are then aggregated. The

hyperparameters that were tuned in this model were the number of estimators, how many decision tree stumps were performed and aggregated, and the learning rate. The hyperparameters chosen from the grid search were a learning rate of 1.5 and 500 for the number of estimators.

Gradient Boosting is an ensemble method where decision trees are also used. Unlike Adaptive Boosting, Gradient Boosting uses decision trees with a greater depth than one. Each Decision Tree is fitted on the errors of the previous, then the results are multiplied by the learning rate. The hyperparameters chosen from the gird search were a learning rate of 0.1 and 500 for the number of estimators.

Naïve Bayes was not chosen due to the mixture of both categorical and continuous data in the dataset. A Decision Tree model was not chosen as several ensemble methods, which use multiple Decision Tree models were used. These often perform better than a single Decision Tree model which can be prone to overfitting, however they do reduce interpretability of the model.

Model evaluation

The best performing models from the hyperparameter tuning were then tested using a ten-fold cross validation approach. Figure 4 and Table 3 show a comparison of the accuracy scores of the models. The standard deviation of the accuracy scores is shown in Table 3. Many of the models performed well with an accuracy score over 78%. The adaptive boosting model performed poorly compared to the other models with an accuracy score of 45%. The best performing model was the Random Forest model with an accuracy score of 86%, followed by the Gradient Boosting model with an accuracy score of 83%.

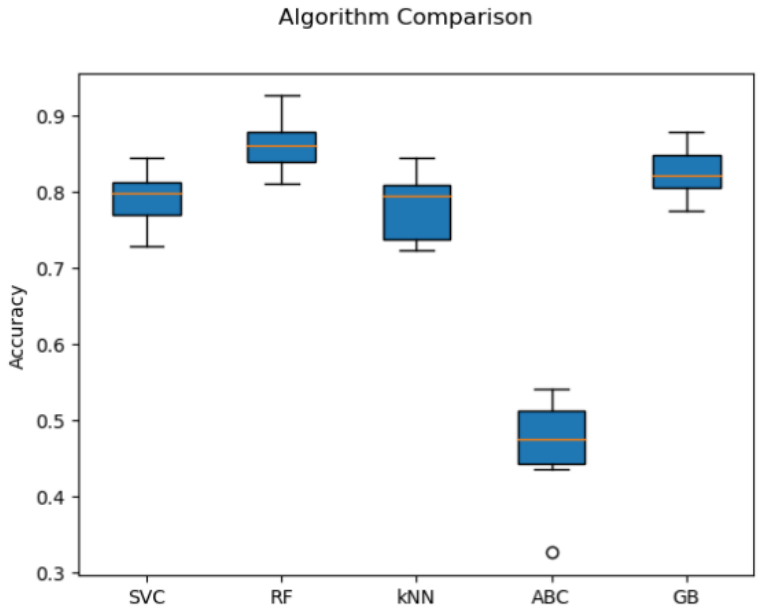

Figure 4. Box plot showing the accuracy scores produced from the 10-fold cross validation of each trained and test model.

Model	Accuracy	Standard Deviation
SVC	0.788	0.035
RF	0.855	0.031
KNN	0.781	0.044
ABC	0.448	0.060
GB	0.830	0.027

Table 3. The accuracy score and standard deviation for each trained model.

The validation set was used to show the out-of-sample performance of the final Random Forest model. Table 4 shows the accuracy, weighted F1 score and the area under curve receiver operator characteristic (AUC-ROC) for the random forest model with the validation set. These scores show that the Random Forest model performs well. These scores show that the model is not likely to be overfitting the data as the model shows similar accuracy performance on the training and validation set.

Accuracy	Weighted F1	AUC ROC
0.826	0.830	0.974

Table 4. The accuracy, weighted F1 and AUC ROC scores for the Random Forest model.

Figure 5 shows the importance of each feature is in the Random Forest model. Figure 5 shows that age is the most important feature, followed by frequency of vegetables eaten (FCVC), number of main meals (NCP), Time using Technology (TUE), consumption of water daily (CH2O) and physical activity frequency (FAF). To further develop the model, feature engineering could be performed. For example, selecting features based on a feature importance threshold to be included in the model, and comparing the performance of the model with all the feature to the model with the engineered features.

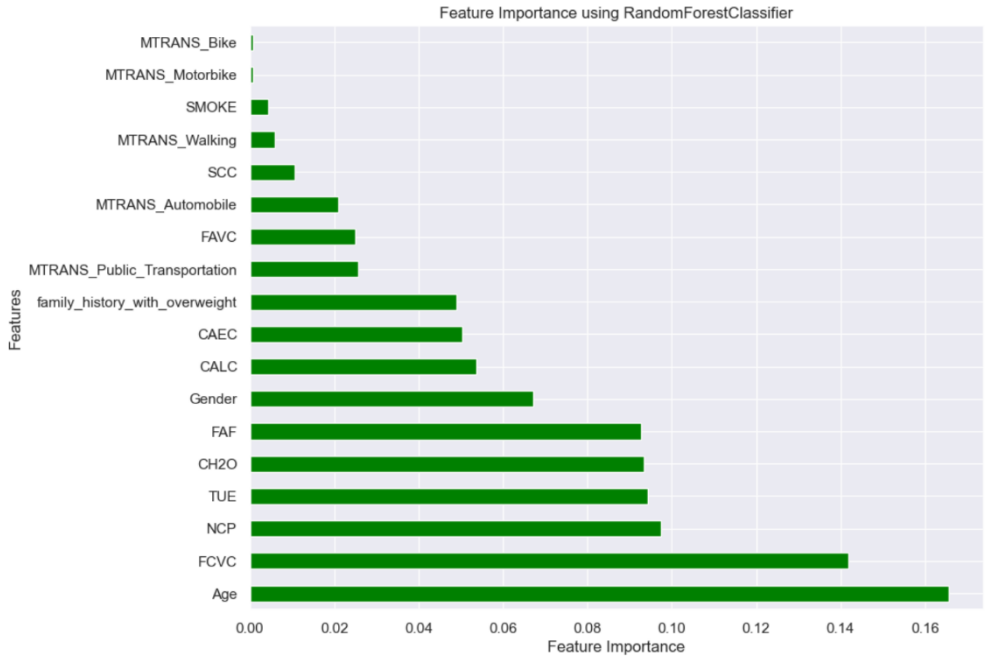

Figure 5. The feature importances of each feature in the tuned Random Forest Model.

Limitations and future work

This model uses classification algorithms to predict the obesity category of an individual from lifestyle factors. The category for the obesity level is determined by the BMI of the individual (equation 1). Another approach to this problem, which could use the same data, would be to perform a regression model to predict the BMI of an individual from their lifestyle factors. The performance of these regression models could then be compared to the performance of the classification models, this would give a greater number of possible models to compare the performance.

A limitation of this model is that it only considers a few features which contribute to obesity. This model only looks at lifestyle factors for prediction. Obesity is a disease that has many factors that contribute to its development in individuals. Future work could look at including a more diverse range of features, such as including genetic and socioeconomic factors into the predictive model. This would lead to a complex model with many dimensions. Dimensionality reduction techniques would likely be needed in the development of this model.

A further use for predictive models regarding obesity and preventative healthcare could be to build a model to predict the obesity related health conditions an individual may be susceptible to. For example, predict whether an individual with obesity risk factors is more susceptible to developing diabetes or a cardiovascular disease. This would enable targeted treatment to the individual to help them prevent the development of the that disease, this would reduce the load on the health care system in the long run.

Andreyeva, T., Sturm, R. and Ringel, J.S. (2004) Moderate and severe obesity have large differences in health care costs. *Obesity Research* [online]. 12 (12), pp. 1936–1943.

Bischoff, S.C. *et al.* (2017) Towards a multidisciplinary approach to understand and manage obesity and related diseases. *Clinical Nutrition* [online]. 36 (4), pp. 917–938. Available from: https://www.sciencedirect.com/science/article/pii/S0261561416313231 [Accessed 23 August 2023].

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* [online]. 16, pp. 321–357. Available from: https://www.jair.org/index.php/jair/article/view/10302 [Accessed 23 August 2023].

Eckel, R.H., Kahn, S.E., Ferrannini, E., Goldfine, A.B., Nathan, D.M., Schwartz, M.W., Smith, R.J. and Smith, S.R. (2011) Obesity and Type 2 Diabetes: What Can Be Unified and What Needs to Be Individualized? *The Journal of Clinical Endocrinology and Metabolism* [online]. 96 (6), pp. 1654–1663. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3206399/ [Accessed 23 August 2023].

Kinlen, D., Cody, D. and O'Shea, D. (2018) Complications of obesity. *QJM: An International Journal of Medicine* [online]. 111 (7), pp. 437–443. Available from: https://doi.org/10.1093/qjmed/hcx152 [Accessed 23 August 2023].

Palechor, F.M. and Manotas, A. de la H. (2019) Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief* [online]. 25, p. 104344. Available from: https://www.sciencedirect.com/science/article/pii/S2352340919306985 [Accessed 23 August 2023].

Prospective Studies Collaboration *et al.* (2009) Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. *Lancet (London, England)* [online]. 373 (9669), pp. 1083–1096.

Safaei, M., Sundararajan, E.A., Driss, M., Boulila, W. and 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* [online]. 136, p. 104754. Available from: https://www.sciencedirect.com/science/article/pii/S0010482521005485 [Accessed 23 August 2023].

Tremmel, M., Gerdtham, U.-G., Nilsson, P.M. and Saha, S. (2017) Economic Burden of Obesity: A Systematic Literature Review. *International Journal of Environmental Research and Public Health* [online]. 14 (4), p. 435. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5409636/ [Accessed 23 August 2023].

Uçar, M.K., Uçar, Z., Köksal, F. and Daldal, N. (2021) Estimation of body fat percentage using hybrid machine learning algorithms. *Measurement* [online]. 167, p. 108173. Available from: https://www.sciencedirect.com/science/article/pii/S0263224120307119 [Accessed 23 August 2023].

Williams, E.P., Mesidor, M., Winters, K., Dubbert, P.M. and Wyatt, S.B. (2015) Overweight and Obesity: Prevalence, Consequences, and Causes of a Growing Public Health Problem. *Current Obesity Reports* [online]. 4 (3), pp. 363–370. Available from: https://doi.org/10.1007/s13679-015-0169-4 [Accessed 23 August 2023].