Obesity Classification with Random Forests and Naïve Bayes

Jakob Brown | City University of London | School of Mathematics, Computer Science, and Engineering

Problem Motivation and Description

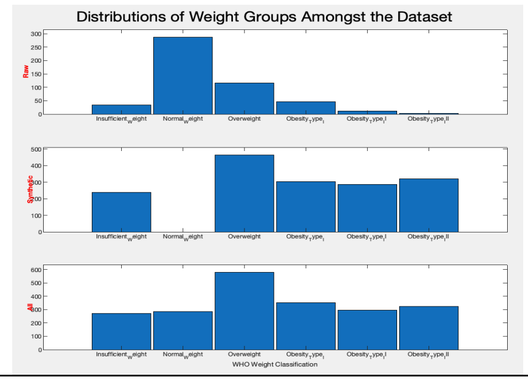
- The pervasion of Obesity across the globe is estimated to have tripled since 1975. The disease and its ramifications include impaired quality of of life, as well as increased mortality through association with myriad health conditions including cardiovascular diseases and cancers[1, 3]
- Obesity is on the rise, however it is preventable: increased knowledge of risk factors and health effects, and increased levels of data collection presents the opportunity to predict and diagnose the disease, from which the epidemic can be addressed effectively[3, 4]
- The objective of the present work is to contribute to the body of research that is employing machine learning methods in order to effectively diagnose and classify obesity levels by making use of data containing identified risk factors for obesity.
- Furthermore, the author hopes to build on the results obtained by De-La-Hoz-Correa et al (2019) by achieving comparable classification efficacy while omitting the synthetic data produced and used alongside the authentic data collected in the aforementioned study, in the hope of producing more generalizable models for future real-world data.[5, 6]
- In the process, Random Forest (RF) and Naïve Bayes (NB) classifiers will be applied to the problem, followed by critical evaluation of the two methods and comparison with other peer-reviewed results in the domain, including the reference paper.[5, 6,]

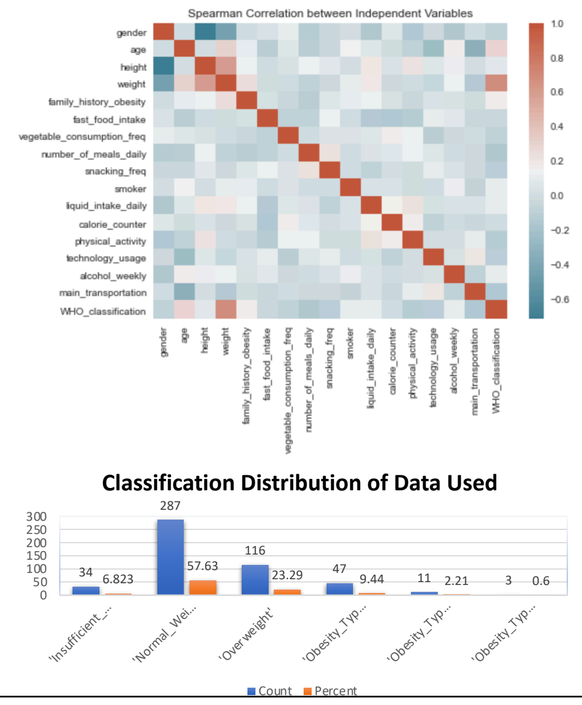
Preliminary Analysis of Data Set inc. basic Statistics

- The data set: 'Estimation of obesity levels based on eating habits and physical condition Data Set' from UCI Repository.
- The data set contains 16 predictor variables, 3 numeric (ratio) and 13 categorical, regarding the physical condition and lifestyle habits of individuals from Mexico, Colombia, and Peru (*see summary statistics).*
- The Data was checked for collinearity between independent variables using Spearman's rank correlation, which of which there was very little (*see heatmap*). Correlations with the dependent variable (WHO_classification) initially seem scarce, save for weight of course, and some moderate relationships between age, physical activity, and number of meals daily. Although this is not ideal, it suggests an opportunity to reduce dimensionality, which in general can benefit predictive capacity through simplification of data input and greater consequent generalisability to unseen data. This will be explored in the present analysis.
- The target variable is made up of 7 classes of weight, which are based on the obesity classification employed by the World Health Organisation[6]
- To allow comparability with the reference paper results, two target variable classes ('Overweight_level_I' and 'Overweight_Level_II') were merged to make 6 classes.[5]
- The original data set contained 2,111 points, 77% percent of which were synthetically generated by the WEKA tool and SMOTE filter. This was done in order to address the uneven class distribution observed in the raw data, collected on an online survey platform (*see Distributions of Weight Groups Amongst the Dataset*).
- Several categorical variables were numbered, allowing the synthetic tools to generate data that lay 'in-between' categories e.g. values of 1.2837 despite categories containing integers ranging from 1-4.
- Moreover, initial modelling using basic decision tree and NB methods on the whole data set quickly replicated results produced by De-La-Hoz et al (2019)[5,] with accuracy of 97%+.
- The present author decided to omit this synthetic data from the current study, on the grounds that it did not fully represent the original content of the features being recorded, and so ran the risk of over-fitting to the data set at hand. This also gave rationale for implementing RF as a method: could either of the two methods, combined with appropriate hyperparameter tuning, feature selection, and/or engineering produce credible results without using the synthetic portion of the data set? And could they produce results as successful as the decision tree and NB methods employed in the reference paper, in spite of decision trees tending to produce better results with smaller data sets.[7]

Summary Statistics

Feature	Age	Height	Weight	Family history of Obesity	Snacking Frequency	WHO classification
Count	498	498	498	498	498	498
Mean	23.14	1.68	69.57	0	0	0
Std	6.72	0.09	17.01	0	0	0
Skew	2.21	0.2	1.19	0	0	0
Top (cat)	0	0	0	Yes	Sometimes	Normal Weight
Freq (cat)	0	0	0	300	289	287





Naïve Bayes

- A probabilistic model based on Bayes Theorem. Class membership is determined by calculating the product of the probabilities of each observation belonging to a each class given the values of each predictor variable for that observation, and then attributing the observation to the class with the greatest probability.
- Central to the function of Naïve Bayes is the assumption of independence of every feature input to the classifier. While this is very rarely true (hence the name *Naïve* Bayes),.

Pros

- ✓ Works well with large, wide datasets with lots of dimensions e.g. text[10]
- ✓ Is quick and easily scalable as a method, with the maximum-likelihood function meaning evaluation of an expression is done in linear time, and the number of parameters required is linear with the number of variables
- ✓ Has a high degree of interpretability, given the simplicity of its underlying principles
- ✓ Multiple different probability distributions can be applied to the data, such as gaussian, Bernoulli, multivariate multinomial etc., making it relatively flexible as a method.

Cons

- ○ If the assumption of independence is confounded, this can often lead to poor performance.
- ○ If a class is not represented in the training set, but is encountered in the test set, Naïve Bayes will attribute a zero probability (known as zero frequency). This is treated with a smoothing function, which could be considered an over-head
- ○ Typically does not perform as well as other models, and thus is used more as a popular baseline method.[9]

Random Forest

- A number of decision trees, operating as an ensemble, used for both classification and regression. Each is trained on a 'bagged' subset of the training data.
- As a classifier, the individual trees will be created and give a predicted class for each observation, which will then be averaged, meaning the class with the 'majority vote' for any given observation will be selected and attributed to it.

Pros

- ✓ Complements the shortfalls of individual decision trees by working as an ensemble: so long as the individual models are uncorrelated enough, the individual errors of each tree will be caught and filtered out through the majority vote process, thereby preventing overfitting.
- ✓ Is capable of ranking features by their importance, which can be a highly useful trait for feature selection or gaining further insight.
- ✓ The ensemble nature allows RF to perform better than other methods on imbalanced data (pertinent to the present data), or data with outliers or missing values.

Cons

- ○ When performance is increased via more sophisticated methods, there is an inevitable trade-off on interpretability.
- ○ The features input into the model need to have some predictive power, else they may hinder the performance of the model as a whole. [11]
- ○ Error can slip through the filtering process of Random Forest's majority vote classification if the trees that make up the ensemble are too correlated/similar.

Hypothesis Statement

- It is posited that the results of both methods will produce worthwhile results.
- RF is hypothesized to perform better than NB, on the basis of RF generally performing better than Naïve Bayes in other peer-reviewed classification tasks in the wider literature[10, 11, 12], and the results in the reference paper, which showed a slight edge in performance of decision tree-based modelling compared to NB, which should generalize to a RF ensemble. NB, as a high bias, low variance classifier, does have an advantage with smaller data sets like the one being used in that it is less likely to overfit and yield poorer test performance.
- Furthermore, in relation to the reference paper, results of the present study can be expected to be at least slightly lower than those observed by De-La-Hoz Correa et al (2019), given the decision to omit the synthetically generated data.

Training and Evaluation Methodology

- 3 different forms of the data set were produced in pre-processing: The typical categorical-heavy data set previously described; a version of the data in which the categories were integer encoded to be treated as numeric; and finally a version in which all categorical variables were one-hot encoded to dummy variables. NB and RF models were optimized, trained, and subsequently evaluated on all 3 versions of the data.
- Feature selection/dimensionality reduction, PCA, and some feature engineering were also tested with each version of the data.
- Said results were measured by taking 10-fold cross-validation across 80% of the data in the model selection stage, giving a proxy for indicating the generalizability for the model. In line with the reference paper[5], precision and recall were used as metrics on a 20% test set which were used in lieu of accuracy on account of the imbalance in the target variable's distribution. F1 scores were calculated from these metrics both on the present results and on those of De-La-Hoz et al(2019) in order to give a more balanced comparison of performance. Cross-validation error was preferred to out-of-bag error for RF as it allowed fair comparability with NB.
- The final models ultimately presented below were trained on a dummy variable version of the data, with BMI engineered as a feature from height and weight predictors, which was then inserted in place of them.

Naïve Bayes Parameters

- Employed a gridsearch algorithm to ascertain the optimum hyperparameters, which were as follows:
  - ○ Distribution name: Kernel
  - ○ Width: 0.51819
  - ○ Kernel: Epanechnikov
- The 'Normal' distribution was not applicable to the present data, due to there being zero variance between certain dummy variables and certain target classes.

Naives Bayes results

- Performed better when working with dummy variables compared to numerically encoded or categorical equivalents.
- Performance was also greatly dependent on feature reduction: including less important variables at the training stage resulted in weaker performance.
- Near-perfect accuracy and F1 scores could be generated with NB, just as with RF, through feature engineering and reduction of the data. The model selected was the best performing that still allowed some comparison with RF.

Choice of parameters and Experimental Results

Naïve Bayes		Random Forest
94.22%	Training accuracy	99.25%
94.72%	Cross-validation accuracy	99.25%
0.74	Test F1 score	0.81

Random Forest Parameters

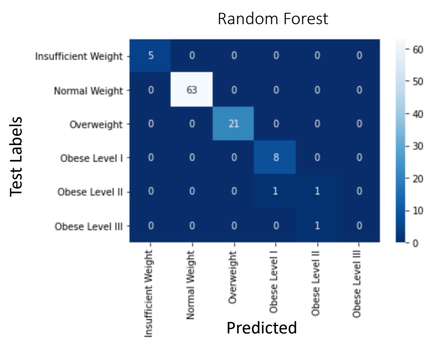
- Used the bagging method to group together weak tree classifiers
- Employed a gridsearch algorithm to ascertain the optimum hyperparameters, which were as follows:
  - ○ Number of learning cycles: 136
  - ○ Minimum leaf size: 2
  - ○ Maximum number of splits: 4
  - ○ Split Criterion: gdi
  - ○ Number of variables to sample: 37

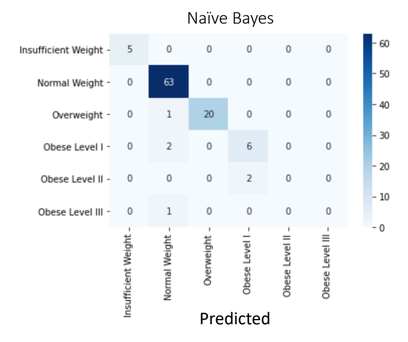
Random Forest Results

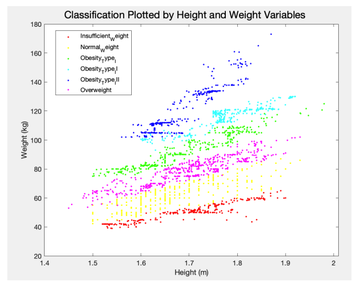
- RF consistently performed similarly across the different types of data, as decision trees' structure are well suited to multi-class categorical data.[16]
- Optimization run-time was longer than NB (c. 55 minutes vs c. 6 minutes for NB), as was training time of the individual model (5.47 seconds vs 2.109 seconds).
- Optimized models consistently returned stronger performance measures compared to NB.

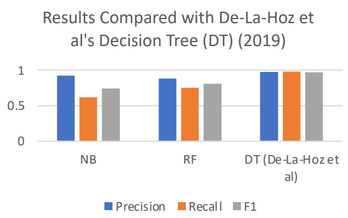
- Analysis and Critical Evaluation of Results

- The Hypothesis statements can generally be accepted in light of the results of the optimized NB and RF models presented, with the exception of the 3rd, which posited that the present models would not achieve as good results as those of De-La-Hoz et al (2019). While this is the case for the models presented, further feature reduction combined with the feature engineering was capable of yielding results akin to those of the reference study (see '*The Effect of Feature Engineering and reduction on results*').
- The presented models were chosen as they gave a balance between producing relatively decent results, while still giving some difference between the two methods in order to be able to provide critical analysis.
- The near-perfect results akin to the reference paper can be attributed heavily to the BMI feature engineering, which, through modelling with fewer and fewer features alongside it, can be seen to have a direct correlation with the target variable (see '*Classification Plotted by Height and Weight Variables*'). This significantly boosted the performance of both methods.
- While Naïve Bayes's high bias can allow its output to swayed by features that are not particularly good predictors, thus resulting in underfitting, random forest is able to achieve low bias through its random feature selection, while also retaining relatively low variance through its robust nature in randomly bagging different subsets of the data.
- Poorer NB performance may be due – in part - to the confounding of the assumption of independence between variables.
- The random selection of features for each weak learner in the RF ensemble proved to be a key different in performance: this allows the RF to discern signal from noise extremely well, hence the less drastic improvement in F1 score across data sets with different amounts of predictors compared to NB.
- One factor that appeared to be a hindrance on the performance of both methods was the imbalanced structure of the data's classes.
- In instances in which predictions are incorrect, NB can be seen to more frequently allocate those predictions to the dominant 'Normal weight' category, again indicating the inherent bias in its algorithm, which leads it to perform worse on imbalanced data sets such as the present one[9].
- While RF often made its mistakes by classifying one label adjacent (which can be interpreted as 'closer' given the ordinal nature of the classes), these misclassifications were entirely made with the much less frequent classes of the higher levels of obesity, the most extreme of which there were only two training examples and one test example.
- While the different versions of data (categorical, integer encoded, and one-hot encoded) virtually made no different in performance for RF, NB performed slightly better with the one-hot encoded dummy variables bolted on to the numeric ones. This reflects NBs preference for data that is not mixed In type, as well as the binary nature of dummy variables lending themselves well to probabilistic methods, as they struggle to interpret ordinality[16].
- Across fluctuation of hyperparameters for both methods, a tradeoff of precision and recall could be observed but could not be avoided. The scarce amount of some of the classes in the training and test sets meant that more often, recall was low, and precision was high, because the models would not predict the rarer classes, and so the completeness of positive class retrieval was hindered while the purity of positive class retrieval overall was not. For this same reason, even with the modelling of BMI as a sole predictor against the target variable, recall never rose above 0.9, while precision saw levels of up to 0.9778.
- RF handled this tradeoff better, with a consistently smaller gap between precision and recall than NB, resulting in a greater overall F1 score, even when NB occasionally exceeded RF in precision.
- For models to successfully classify these evasive classes, the data would need to contain more of them, which is likely why the reference authors synthetically augmented the data set to bolster the more uncommon naturally-occurring classes of obesity. The benefits of this is highlighted by the high levels of both precision and recall in the research conducted by De-La-Hoz et al (2019).
- Finally, one way in which NB does appeal more than RF is computational exertion and interpretability. For the present study, neither are too problematic, however for time or resource-sensitive problems which require explaining, RF may prove inapplicable.









The Effect of Feature Engineering and reduction on results

	Height & weight + all predictors	Height & weight + significant predictors	Height & weight	BMI + all predictors	BMI + significant predictors	BMI
RF F1 score	0.74	0.77	0.78	0.81	0.94	0.94
NB F1 score	0.48	0.48	0.49	0.74	0.86	0.94

Lessons Learned and Future Work

- Random Forests do indeed tend to perform well with any type of data, due to their robust ensemble nature.
- Naïve Bayes needs more help in data preparation and optimization to produce comparable results with RF.
- Future work on RF: Investigate pruning and tree depth as another hyperparameter, and how this interacts with the reduction of features to see if there is an-optimum trade-off as detail and nuance is reduced by both of these factors. Also investigating other tree ensemble aggregation algorithms, for example AdaBoost, which returned excellent results in preliminary experimental modelling on the present data.
- Future work on NB: Investigate the effects of standardization of the data on performance.
- Future work for both: Explore the benefits of stratified sampling for both train and test sets.
- While the methods outlined provide an effective tool in predicting obesity level from various lifestyle and physical predictors, its usefulness only extends as far as being able to identify the direct relationship with which height and weight have on the target variable which is classed by BMI rating[5, 6]. Future work should seek to create diagnostic tools from purely indirect indicators of obesity. This, as well as work with longitudinal data which allows us to see how successfully a model can predict obesity in the future before individuals actually develop the condition, will give invaluable insight into the risk factors and conditions that lead to the development of the disease[4, 13]. From this knowledge, powerful preventative strategies can be devised and implemented at a global level, and treatment and prevention can also be tailored more confidently at an individual level[4, 8, 13].

References

1. Abdelaal, M., le Roux, C.W. and Docherty, N.G., 2017. Morbidity and mortality associated with obesity. *Annals of translational medicine*, *5*(7).
2. Rivera, J.A., Barquera, S., Campirano, F., Campos, I., Safdie, M. and Tovar, V., 2002. Epidemiological and nutritional transition in Mexico: rapid increase of non-communicable chronic diseases and obesity. *Public health nutrition*, *5*(1a), pp.113-122.
3. Haslam DW, James WP. Obesity. Lancet 2005;366:1197-209. 10.1016/S0140-6736(05)67483-1
4. Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I. and Keane, J., 2009. Comparing data mining methods with logistic regression in childhood obesity prediction. *Information Systems Frontiers*, *11*(4), pp.449-460.
5. De-La-Hoz-Correa, E., Mendoza Palechor, F., De-La-Hoz-Manotas, A., Morales Ortega, R. and Sánchez Hernández, A.B., 2019. Obesity level estimation software based on decision Trees.
6. Palechor, F.M. and de la Hoz Manotas, A., 2019. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in brief*, *25*, p.104344.
7. Ali, J., Khan, R., Ahmad, N. and Maqsood, I., 2012. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), p.272.
8. Cervantes, R.C. and Palacio, U.M., 2020. Estimation of Obesity levels based on computational intelligence. *Informatics in Medicine Unlocked*, p.100472.
9. Lewis, D.D., 1998, April. Naive (Bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning* (pp. 4-15). Springer, Berlin, Heidelberg.
10. Pranckevičius, T. and Marcinkevičius, V., 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), p.221.
11. Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R news*, *2*(3), pp.18-22.
12. Caruana, R. and Niculescu-Mizil, A., 2006, June. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).
13. Dugan, T.M., Mukhopadhyay, S., Carroll, A. and Downs, S., 2015. Machine learning techniques for prediction of early childhood obesity. *Applied clinical informatics*, *6*(3), p.506.
14. Adnan, M.H.B.M. and Husain, W., 2012, June. A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction. In *2012 International Conference on Computer & Information Science (ICCIS)* (Vol. 1, pp. 281-285). IEEE.
15. Breiman, L., 1999. Random forests. *UC Berkeley TR567*.
16. Singh, A. and Lakshmiganthan, R., 2018. Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms.