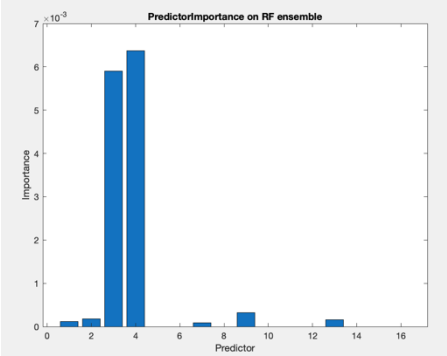
Supplementary Materials

Glossary

Bagging – One of the methods of selecting sub-sets of the training data with which to grow individual tree learners in random forests. Bagging randomly selects data points without replacement.
Bayes Theorem – A theory of probability and statistics, which describes the process of calculating the probability of an event based on some known prior conditions or probability that play a part in determining the probability of the event in question.
Collinearity – refers to the correlation between predictors (or independent variables), such that they share a linear relationship between them.
Cross-Validation – the process of splitting a data set in to *k* folds, and subsequently training a machine learning model *k* times. Each subset of the data is used as a holdout test set once, and thus used to train the model at hand *k*-1 times. The training accuracy
Decision Tree – With regards to machine learning, a decision tree is a form of learner, which can be used in classification or regression. It is based on a tree-like structure, made up of nodes which split on values of given variables in a data set as a data point descends down the tree, before finally reaching a terminal node, which either will classify the data point or average its numerical value.
F1 score – This measure refers to the figure between 0 and 1 that results from a balancing equation performed on both precision and recall measures (see below). The F1 measure therefore reflects both precision and recall performance of a classification model in one metric.
Maximum likelihood – a particular way of approximating the parameters of a probability distribution, which is seen in many probabilistic methods of machine learning, including Naïve Bayes.
Overfitting – The process of creating a machine learning model, or any analysis, that too closely represents a particular set of data, and consequently does not represent other sets of data as well, nor the greater population of data from which the particular subset has been drawn.
PCA – Principal Component Analysis: a dimensionality-reduction technique, which performs a linear transformation numeric data points (treated as vectors).
Precision - In classification, this refers to the quantification of the number of positive class predictions returned that are actually belonging to the positive class. In cases of multi-class classification such as this, precision is calculated for each of the possible classes, and is then averaged by the mean.
Recall – In classification, this refers to the quantification of the number of positive class predictions returned out of all examples of that positive class in the data set. In cases of multi-class classification such as this, recall is calculated for each of the possible classes, and is then averaged by the mean.
Smoothing function – An optional factor in the Naïve Bayes algorithm, often applied, which adds a small scalar to each of the inputs of the probability equation, to ensure no zero-values result in a posterior probability of zero after their product has been calculated.
Spearman's Rank Correlation – a non-parametric statistical method to calculate the rank correlation between two variables or sets of data

Intermediate Results

Discussed in the poster is the discovery that the predictive power of the data was mainly located in two variables height and weight, which on their own could predict almost as well as the entire data set, and when combined together to make BMI, could predict at a near-perfect rate. Below is the predictor importance bar plot showing the imbalance in signal that was picked up by the RF ensembles' 'PredictorImportance' function. This was deemed a negative result, as it made it difficult to derive interesting interactions of the data, and detracted from the tuning of the models themselves, as performance so strongly depended on these features. In hindsight, the data set was not the best for the assignment of recording, comparing, and critically evaluating two machine learning model's performance. This however was discovered after completing pre-processing in MATLAB, by which time it would have been a rush to have gone back and chosen another data set. I thus elected to present not the two technically best performing models in the poster, but rather two models that had some difference and imperfections that allowed critical evaluation. Then, I wrote about the caveat and reported the better results from the BMI-focussed models in a table on the poster. I hope my reasons for presenting the results the way I did can be understood.



It was deliberated and experimented with to try to 'break' the classification problem, by occluding one of the two height and weight variables, preventing the models from drawing a linear relationship between them as a pair and the target variable of BMI. This however could not yield any credible results, meaning that they had to be included, and so the rationale of attempting to produce near-perfect results like that of the reference paper was selected, even if that meant that the findings were not as interesting or impressive. A middle ground was found in the version of the data that incorporated the BMI feature engineering, but did not use any feature reduction: this gave reasonably disparate results between the naïve Bayes and Random forest models, while still achieving decent results for both. The near-perfect results of De-La-Hoz et al (2019) could be replicated simply by reducing the features, as reported in the post-processing table in the poster.

Not included in the results were that of PCA, which were found to produce slightly weaker results for both models when performed on both the integer encoded and one-hot encoded versions of the data set. This probably reflected the slight variance lost in the sub-setting of the 2 greatest principal components, which accounted for an average of 98.88% of the variance.

Other methods for producing and tuning models that were trained are included in the *_model_selection.m files in the zip folder. This includes training models on many variations of the data set, as well as many different ranges and many different sets of hyperparameters.