An Ensemble-Based Feature Selection Method – A Case Study of Childhood Obesity

Abstract. The increasing prevalence of childhood obesity makes it essential to study the risk factors with a sample representative of the population covering more health topics for better preventive policies and interventions. Therefore, a new feature selection method for large-scale and high-dimensional data with good interpretability is proposed in this paper, which is a Bagging-based Feature Selection method integrating the MapReduce technique (BFSMR). The model comprises a collection of feature selection models from filter, wrapper, and embedded feature selection methods and aggregates the selected features from the five methods by voting with both feature weights and model weights. The detailed experiment demonstrates that the method is relevant to real world applications with data collected from the Basque Country (Spain) covering various health topics from primary care, specialty care and nurse-collected questionnaires, including maternal information to explore hereditary factors. The final results show the advantage of avoiding model bias and providing a more reasonable selection of features of higher clinical relevance and interpretability.

Keywords: Feature Selection, Ensemble Learning, Childhood Obesity.

1 Introduction

Childhood obesity has emerged as an important public health problem countries in Europe and the rest of the world. According to the WHO [30], the worldwide prevalence of obesity nearly doubled between 1980 and 2008 and one in three 11-year-old children is overweight or obese in Europe. A childhood obesity review has shown that the increasing prevalence of childhood obesity is associated with the emergence of comorbidities previously considered as "adult" diseases such as diabetes and hypertension which can track into adulthood [22]. Therefore, it is essential to study the risk factors of childhood obesity to design preventive policies or interventions and to evaluate the efficacy of the interventions in place. This should include not only those from the health system but from other areas as well, such as changes in school menus and promotion of physical activity.

The increasing prevalence of childhood obesity is the consequence of an interaction among a complex set of factors that are related to the environment, genetics, and ecological effects such as family, community, and school [22]. There are some surveys capturing the reality of overweight and obese children, but these surveys covered a limited sample and a limited time granularity. More comprehensive data about childhood obesity with a larger sample representative of the population is necessary. Moreover, the interventions need to be established within general practice, reported in the scientific state of art, or supported by experts' experience and knowledge. With more comprehensive data, it is not clear which factors are responsible or affect the process

of a child evolving to the overweight or obese state. In this case, intervention areas are not clear. Data analysis can select the most important features grounded in real data, which can reduce the cost of testing new factors and accelerate the establishment of policies or interventions that are more "evidence-based". From a technical perspective, data analysis comprises multiple tasks throughout the whole process, including data collection, data cleansing, data pre-processing, data modelling and so on. Machine learning techniques are widely applied to accomplish these tasks, providing decision support to the policy-makers.

To reduce the features of the data, two main classes of machine learning models can be applied, namely feature extraction methods and feature selection methods. The difference between the two types of models is that feature selection methods keep a subset of the original variables while feature extraction methods combine the original variables into a smaller set of new features, such as principal component analysis (PCA), linear discriminant analysis (LDA) and Autoencoder. In our study, we focus on feature selection methods to preserve the semantics of the features, as the results need to be interpretable without a subjective definition of new features. Moreover, it is easier for clinicians and policy-makers to establish follow-up interventions for a single feature than a compound factor of multiple features.

There are three main types of feature selection methods: filter methods, wrapper methods, and embedded methods [1]. Filter methods select features based on a statistical measure to assign a score to each variable to rank the variable importance regardless of the model. These methods are time-efficient and robust to overfitting, but tend to ignore the possible interactions between variables. Wrapper methods have the opposite advantages and disadvantages by converting the feature selection task into a search problem. Subsets of variables are compared with other subsets to select the group of features that can give the best predictive performance. Embedded methods are learning models that can perform feature selection and classification simultaneously by integrating feature selection algorithm as part of the learning process. The embedded methods take variable interactions into consideration and are less computationally demanding than wrapper methods. However, in some cases, the optimal feature set selected by one embedded method is classifier-dependent, meaning that the optimal set only works for this specific classifier and cannot contribute to good prediction when used for other embedded classifiers [16], as the optimal set is based on the hypotheses of the classifier.

The advantages and disadvantages reveal that different feature selection methods have their preference when selecting variables and the results may be biased because of the model limitations if we only rely on one method. This problem is even more critical in exploratory research when the problem is not clear and the validity and credibility of the results are crucial. In addition, most wrapper methods and embedded methods select features based on prediction performance, which may lead to a selection of variables with no clinical relevance that are difficult to interpret. Finally, time efficiency will be more challenging when using a large-scale sample representative of the population with a high dimensional structure.

To overcome these limitations and solve the problems in a real-world setting, we propose BFSMR (Bagging-based Feature Selection method integrating MapReduce), a novel machine learning method that can perform efficiently with large-scale data and

combine the results from different feature selection methods to give a more convincing and interpretable selection of features. Moreover, during the process of calculation using various feature selection methods, we can get a comprehensive understanding of the potential important risk factors preferred by different models.

2 Methods

The BFSMR method proposed for the first time in this paper is a bagging-based feature selection method integrating MapReduce, which is a novel method with a good balance between valid results and good interpretation. In this section, we first introduce the MapReduce technique, the bagging method, and the feature selection models used in the bagging framework. Then we construct a new framework incorporating the advantages of MapReduce and bagging at the same time.

2.1 MapReduce

MapReduce is a method to process and generate large-scale data in a parallel and distributed way, which can be very useful in the context of feature selection when the dataset is large-scale and high-dimensional. After splitting the input data into smaller subsets, the model extracts the information of interest in each subset and then merges them to output the aggregated results, ensuring to process large-scale data rapidly. The whole procedure can be broken down into two main tasks, Map and Reduce [8]. The original data is split into an appropriate size and each split is assigned with one Map function defined with respect to data structured in (key, value) pairs. The Map function works in parallel to convert every pair in the input data, denoted as (*k1, v1*), into a list of pairs in a different data domain, denoted as (*k2, v2*). Next, all the pairs (*k2, v2*) with the same key are collected to form one group for one key. Then the Reduce function is applied to each group in parallel and the collection of all Reduce calls is the final result. The procedures of splitting and mapping makes it possible to process the data in parallel and the procedures of shuffling and reducing merge the information by key variable to reduce the data scale.

$$\mathrm{Map}(\mathrm{k1}, \mathrm{v1}) \rightarrow \mathrm{list}(\mathrm{k2}, \mathrm{v2}) \tag{1}$$

$$\mathrm{Reduce}\big(\mathrm{k2}, \mathrm{list}(\mathrm{v2})\big) \rightarrow \mathrm{list}((\mathrm{k3}, \mathrm{v3})) \tag{2}$$

2.2 Bootstrap Aggregating (Bagging)

Bootstrap aggregating, also called bagging, is an ensemble-learning algorithm that applies different models with different random samples and uses majority voting to combine results for the final decision [2]. The method is incorporated in our model to merge results from different feature selection methods. Given a training data D of size N with correct labels $\omega_l \Omega = \{\omega_1, \ldots, \omega_C\}$ representing C classes, generate T bootstrapped samples D_t of size n by random sampling from D uniformly and with replacement. The model m can be applied with D_t to construct the classifier h_t. With the ensemble $\varepsilon =$

$\{h_1, \dots, h_T\}$ derived from the training process, the unlabeled instance **x** in the testing data is classified into the class that receives the highest total vote.

$$V_j = \sum_{t=1}^{T} v_{t,j}, j = 1, \dots, C \; where \; v_{t,j} = \begin{cases} 1, & if \; h_t \; picks \; class \; \omega_j \\ 0, & otherwise \end{cases} \quad (3)$$

2.3 Feature selection methods

In this study, we select five models as representative methods from filter, wrapper and embedded methods to give a relatively comprehensive discussion and comparison of the feature selection methods. The algorithms are not discussed in detail, only the concepts are briefly introduced to show the characteristics of the methods. Besides, it is not necessary to always use the same feature selection methods. The applied methods can be determined based on the concrete problem and data issues.

Filter methods – Mutual Information (MI). Filter methods use statistical measures to rank the variables. An entropy measure called mutual information (MI) is used to assess the features in this section, which is a measure of mutual dependence between two random variables [5]. To select *k* features is a task of finding a subset of features that have the maximum MI with the target variable. The nearest-neighbor method is used to estimate MI [21].

Wrapper methods – SVM-RFE. Recursive feature elimination (RFE) is a greedy optimization by recursively training the estimator on the pruned set and dropping the features with the least feature importance in each iteration until the desired number of selected features is reached [13]. SVM-RFE is a specialization of RFE, which uses a Support Vector Machine (SVM) as the estimator. We used linear SVM for time efficiency. In addition, it is easier to interpret the variable importance with the use of a linear kernel, as the absolute value of the coefficient indicates the importance for the separation made by the hyperplane.

Lasso regression. Least Absolute Shrinkage and Selection Operator (Lasso) is one type of linear regression with regularization, which is a method for adding additional constraints or penalties to a model [28]. The coefficients of the regression variables are penalized and some of them shrink to a very low level or even zero, which largely reduces the number of variables in the model. The linear regression can be written as $Y = X\beta + \epsilon$ and the estimated coefficients should be the values that can make the loss function at the lowest level. Lasso regression introduces L1 regularization to the loss function of linear regression, meaning that a penalty is added to the loss function as Equation 4. All the data should be standardized before applying the Lasso regression to guarantee that the coefficients are on the same scale. In this case, the absolute value of coefficients can reveal the variable importance.

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right| \quad (4)$$

Ridge regression. Ridge regression is similar to Lasso regression [17]. The difference is Ridge regression uses L2 regularization to the loss function, which enforces the coefficients to be lower but does not enforce them to be zero. The Ridge regression minimizes the impacts of irrelevant factors.

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right|^2 \quad (5)$$

Random forest. A random forest constructs a multitude of decision trees at the training time and outputs the most frequently predicted class of individual decision trees as the final result [3]. Two key concepts give the algorithm the property of randomness. First, each decision tree is built by a bootstrapped random sample drawn from the original training set. Second, instead of the full set of features, only a random subset of features are considered when splitting the nodes in each decision tree.

In our study, we used 50 estimators when training the classifier and used Gini impurity [3] to measure the quality of a split. The most commonly used feature importance is Mean Decrease in Impurity (MDI), however, impurity-based importance is biased towards numeric features or categorical features with high cardinality [26]. To overcome this limitation, we used permutation importance for feature evaluation.

2.4 BFSMR

With all the models and techniques introduced in this section, we propose a model that combines the advantage of MapReduce and Bagging and gives a more reasonable set of selected features with better interpretability. The notations used in this section are listed as below.

Notation	Meaning
$D = \{d_p\}$	D is the input data and is split into P chunks, denoted as $d_p,\ p = 1, \dots, P$
c_i	Feature selection classifier where $i = 1, \dots, M$
cid	The Classifier ID where $cid \in \{1, \dots, M\}$
s	Random sample set with Set ID $sid \in \{0, 1, \dots, M\}$ (sid=0 for test set)
$w1_j$	Feature weights based on the ranking from each classifier where $j = 1, \dots, k$
$w2_i$	Method weights based on the model performance where $i = 1, \dots, M$
f_i	Feature lists derived from M feature selection classifiers, $f_i = \{f_{ij}, j = 1, \dots, k\}$
F	Feature space with unique features from M feature lists, $F = \{f_l', l = 1, \dots, L\}$
V_l	Voting score for each unique feature where $l = 1, \dots, L$

The whole structure of BFSMR is shown in Algorithm 1, appearing almost the same as MapReduce, which splits the input data into chunks and applies Map function and Reduce function. However, the Map function is merged with the bootstrapping procedure in Bagging. For each chunk d_p, given M classifiers to select features, M random samples are drawn with replacement and a set ID sid is assigned to each random set. In addition, a test set is drawn with a sid of 0.

After applying the Map function to all splits, the sets with the same sid are merged to be used as the inputs of the Reduce function. The Reduce function works in parallel

to each group and the original MapReduce method normally applies the same model or function to all groups. But in BFSMR, we match the classifiers with groups based on the *sid* and *cid*, so that the corresponding classifier is applied to different groups, which guarantees the possibility of using different feature selection methods to avoid model bias. The outputs of Reducing are k feature lists selected by M classifiers.

The next step of BFSMR is to merge the outputs from different classifiers using the voting strategy learnt from Bagging method. The voting strategy of Bagging is majority voting with equal probability while we assign weights to each feature in the feature list based on the ranking and assign weights to the classifiers based on their predictive performance. Voting with weights is calculated and the top *k* features with the highest votes are selected as the final results.

Algorithm 1. Bagging-based Feature Selection integrating MapReduce (BFSMR)

1: Split the input data *D* into *P* ChunkData d_p
2: for $p \in \{1, \dots, P\}$ do
3: Apply MAP(Index a_p, ChunkData d_p)
4: Select *M* feature selection classifiers c_i with ClassifierID *cid* assigned
5: for $m \in \{1, \dots, M\}$ do
6: Apply REDUCE(SetID *sid*, Sets [*s1*, *s2*, …])
7: Assign $w1_j$ and $w2_i$ based on ranking and performance, $w_{ij} = w1_j \times w2_i$
8: Get the feature set with unique features *F* from *M* feature lists

$$V_l = \sum_{i=1}^{M} \sum_{j=1}^{k} v_{ij} w_{ij}, where\ v_{ij} = \begin{cases} 1, & if\ f_{ij} = f_l^{'} \\ 0, & otherwise \end{cases}$$

9: Select *k* features with the highest votes

Algorithm 2. Map function and Reduce function

1: function MAP(Index a_p, ChunkData d_p)
2: Split ChunkData d_p into training set *t1* of size *n1* and testing set *t2* of size *n2*
3: for $m \in \{1, \dots, M\}$ do
4: Sample Set *s* of size *n* with replacement from Set *t1*
5: SetID *sid* ← *m*
6: Sample Set *s* of size *n* from Set *t2*, SetID *sid* ← *0*
7: EMIT(SetID *sid*, Set *s*)
8: function REDUCE(SetID *sid*=*m*, Sets [*s1*, *s2*, …])
9: Select Classifier c_i with ClassifierID *cid* = SetID *sid*
10 Train Classifier c_i with Sets [*s1*, *s2*, …]
11: Test Classifier c_i performance on Sets with SetID *sid*=0
12: FeatureList f_i = {*k* features based on feature importance}
13: EMIT(ClassifierID *cid*, FeatureList f_i)

3 Experiments

3.1 Data

Data sources. This study uses three datasets from the databases of the Public Health Provider in the Basque Country (Spain), called Osakidetza[1], which provides services to more than 2,200,000 patients, through 16 hospitals and more than 300 primary health centers. The extraction from Osakidetza was made in the context of the European Union funded H2020 MIDAS[2] (Meaningful Integration of Data Analytics and Services) project. The main objective of the MIDAS project is to provide a unified Big Data platform to address the needs of policy-makers and citizens across Europe by mapping, acquiring, managing, modelling, processing and exploiting existing heterogeneous health care data and other governmental data, together with external open data, to enable the creation of evidence-based actionable information. The Osakidetza dataset was properly anonymized and extracted with the approval of an ethic committee in the Basque Country.

Data information. The Osakidetza dataset from the Basque Country is composed of data extracted from primary care, specialty care and hospital databases, including information about the patient, diagnosis, medical forms, medical appointments, prescription information, and information about the child's mother. The dataset contains the information from around 800,000 children throughout the Basque Country who were under the age of 18 in the year 2000 and includes information on these children until they reached the age of 18.

The entire data set extracted from Osakidetza for the MIDAS project is composed of 9 tables, of which 3 tables have been selected for this study: children's information table, children's forms and the children's mothers' forms. Within Osakidetza, forms are health and lifestyle questionnaires that the GP or the nurse may or may not fill out in a consultation (e.g. height, physical exercise habits, etc.).

3.2 Data Pre-Processing

Data cleansing and harmonization. The datasets extracted from Osakidetza needed cleansing since they had inconsistencies, empty columns, and incorrect formatting values. In addition, data about the forms were originally stored in five different datasets: three about children's forms taken from primary care, specialty care and nurse collection records respectively and two about mothers' forms taken from primary care and specialty care.

First, data harmonization was conducted as different form sources contain variables that refer to the same concept but expressed in different ways and/or in different units (e.g. the "weight" variable was expressed in grams on primary care forms and the "weight" variable was expressed in kilograms on specialty care forms). Second, the five

[1] https://www.osakidetza.euskadi.eus/
[2] http://www.midasproject.eu/

forms datasets were merged into two datasets (i.e. children's forms and mothers' forms) based on childrenID and motherID. Third, the mother's forms data was aggregated by year by using the mean values for the numeric variables and the mode values for the categorical variables and dummy variables with values of 0 and 1. The variables of sex, birthdate, and the child's mother's ID in children's information table were added into children's form data and this merged table was further joined with mother's aggregated forms data based on motherID and registration year in both tables. In this case, the mother's forms data becomes additional variables for the children to provide maternal and hereditary information. Finally, the empty columns were dropped and rows with empty BMI were not useful for our study. In the end, only records of the children under the age of 18 with children's BMI in the range between 10 and 60 were kept.

The final merged data has 1,478,857 records from 426,813 children.

Outcome Indicator. The indicator of childhood obesity was created based on the age- and gender-specific BMI. The definition[3] of childhood obesity is that the child's BMI is higher than the 95th percentile of the age- and gender-specific subgroup in the reference population. However, for the reference population in our study, only the 90th percentile is reported in the statistics[4]. In this case, our outcome is to indicate whether the child has a BMI higher than the 90th percentile of his or her age group.

Data Pre-processing. The outcome was defined based on BMI. In fact, some other variables are used as indicators in obesity-related studies [18], which will show strong correlation with our outcome if these variables are included in the predictors. In this case, 5 variables were dropped, i.e. BMI, height, weight, waist, and size.

There are three types of variables per record: 28 numeric variables, 8 dummy variables, and 33 categorical variables. Almost all variables have many missing values. We replaced missing values with 0 for the numeric variables and with "Missing" for categorical variables. With the refilled "Missing", the *n*-level categorical variables were converted into *n+1* dummy variables by using One-Hot encoding [15]. To overcome the limitation of One-Hot encoding, the dummy of "Missing" was dropped to avoid the multicollinearity [12]. It was then used as the default option for each variable. Eventually, 83 dummy variables replaced 33 categorical variables, with each dummy representing one level of the categorical variables, denoted as "VarName_LevelName". The exception was the "Yes" dummy of the two-level categorical variables, denoted as "VarName". It was tested to process the eight dummy variables in the same way as categorical variables. However, it would lead to severe multicollinearity in the data because of the sparse structure of the dummy variables. In this case, we replaced missing values with 0 for the dummy variables. All the data was standardized before model application.

---

[3] Defining Childhood Obesity https://www.cdc.gov/obesity/childhood/defining.html

[4] https://www.fundacionorbegozo.com/el-instituto-de-investigacion-del-crecimiento-y-desarrollo/graficas-y-tablas/

Table 1. Features used in the experiment.

Numeric Variables (28)	Dummy Variables (8)	Categorical Variables (33/83)
Birthyear,Age,SystolicPressure,DiastolicPressure,Glucose,Triglycerides,UricAcid, Gpt,Got,FruitsPerDay,FruitVegetableConsumption,NumberCigarettes,AlcoholUnitWeek,Birthheight,Birthweight,GestationalAge,CardiovascularRisk, FraminghamCardiovascularRisk, RegicoCardiovascularRisk,MoSystolicPressure,MoDiastolicPressure, MoBMI,MoAlcoholUnitWeek, MoNumberCigarette, MoCardiovascularRisk, MoFraminghamCardiovascularRisk, MoRegicoCardiovascularRisk, MoPhysicalExerciseHours	Sex, DietIntentChange, DietCompliesAdvice, MoRecommendedDiet, MoBirth, MoGestationalDiabetes, MoDietIntentChange, MoUnknownVariable	**YES/NO**(19/38):Alcohol,Diabetes,ExerciseAdvice,BreakfastDairy,BreakfastFruit, DietCorrectExecution, AdequateDietaryKnowledge,MoAlcohol,MoExerciseAdvice, MoFitnessAdvice, MoDiabetes, MoDietCorrectExecution, MoAdequateDietaryKnowledge, MoBirthPreparation,MoBreastfeedingEducation, MoMaternalBreastfeedingInformation, BreastfeedingAbandonment, DietEducation, MoDietEducation **Normal/Abnormal**(1/2): Sleep **Adequate/Inadequate**(4/8): Diet, MoDiet, PhysicalExercise, MoPhysicalExercise **MULTI**(9/45):Tobacco(6),MoTobacco(6), MoSmoker(2), RecommendedDietType(6), MoRecommendedDietType(8), TypeBreastfeeding(4), MoTypeBirth(6), MoPlaceBirth(4), MoPromotionBreastfeeding(3)

Note: All “Mothers-” in the variables were replaced with “Mo-” for shorter names.

3.3 Experimental Setup

The data was imported in chunks with a size of 10,000 rows and we got 148 chunks in total with 147 full chunks and the last chunk only including 8857 rows. The training set and testing set was split at the ratio of 0.8:0.2 and the size of bootstrapped random samples for each feature selection method was 10% of the training set.

The five feature selection methods were applied in parallel and we selected 10 features from each method. We used the nearest-neighbor method to estimate MI in the Filter method and the number of neighbors was 3 [21]. The features that had the maximum MI with the outcome were regarded as features with the highest importance. To avoid the problem of long execution time, linear SVM was applied as the estimator in the SVM-RFE method and the absolute value of the coefficient was the feature importance. The step of the RFE method was set as 1, meaning that one variable was dropped in each iteration and the final 10 variables left in the model were the selected results. To determine the regularization parameter (λ) for Lasso and Ridge regression, the models were iteratively fitted along the regularization path on a grid of parameter, and the parameters that led to the best performance in the cross-validation test were selected, which was 0.002237 for the Lasso regression and 10 for the Ridge regression. As for the Random Forest, it used 50 estimators and permutation importance to select the most important features. All the models were tested on the same testing set. As the

Filter method selected features without learning algorithms, linear regression was applied using the selected features to get the predictive performance.

The model applies the voting strategy that takes both feature weights and method weights into consideration. To examine whether this voting strategy could effectively select the feature set with better interpretability and clinical relevance, three voting strategies were applied and the selected results were compared.

Voting1: Voting with equal score for all features

Voting2: Voting with feature weights

Voting3: Voting with both feature and method weights

4 Results

Table 2 presents the top 10 features selected from different models and the "(-)" beside the variable names means a negative effect on the outcome. There are some common features selected by multiple models, such as Age, Sex, and MothersDietEducation. However, the model preference of different types of variables could still be observed. Lasso and ridge regression were both the specializations of linear regression, with the only difference in regularization method, thus, they selected similar features. The variables selected covered Age, Sex, Birthyear, smoking habits, exercise habits, and diet

Table 2. Top 10 features selected from different models.

	Filter (MI)	SVM-RFE	Ridge	Lasso	RandomForest
1	Age	MoDietEducation	Age	Age	SystolicPressure
2	Sleep_Normal (-)	MoRDType_LowSalt	Sex (-)	Sex (-)	MoDiastolicPressure (-)
3	BFType_Maternal (-)	RDType_2000cal	Tobacco_No (-)	Tobacco_No (-)	MoSystolicPressure (-)
4	DiastolicPressure (-)	AdeDKnowledge	DietEducation	DietEducation	Sex
5	MoSystolicPressure	MoPE_Inadequate (-)	MoTobacco_Yes	MoDietEducation	Birthyear (-)
6	MoNumberCigarettes	DietCompliesAdvice	BFType_Maternal (-)	BFType_Maternal (-)	Tobacco_No (-)
7	Birthheight (-)	MoRDType_Free (-)	PE_Inadequate	Birthyear (-)	MoExerciseAdvice (-)
8	MoBMI	MoPEHour	MoDiabetes_No (-)	MoNumberCigarettes	MoAlcohol_No (-)
9	Birthweight (-)	DiastolicPressure (-)	PE_Adequate(-)	PE_Inadequate	PE_Inadequate
10	MoDiastolicPressure (-)	SystolicPressure	MoDietEducation	DCExecution_No	MoTobacco_Ex

Note: feature abbreviation: RDType - RecommendedDietType, MoRDType – MoRecommendedDitetType, BFType - BreastfeedingType, PE – PhysicalExercise, MoPE – MoPhysicalExercise, MoPEHour – MoPhysicalExerciseHour, AdeDKnowledge – AdequateDietaryKnowledge, DCExecution – DietCorrectExecution,

knowledge. The Filter method showed a preference for numeric variables, such as Age, Birthheight, Birthweights, MothersNumberCigarettes, DiastolicPressure, SystolicPressure, MothersBMI, MothersDiastolicPressure, and MothersSystolicPressure. Moreover, they were also variables with fewer missing values. Random Forest also had an inclination to numeric variables (SystolicPressure, MothersDiastolicPressure, MothersSystolic-Pressure, Birthyear) and variables with fewer missing values (Sex) but it also selected variables about smoking habits, exercise habits, and alcohol use. SVM-RFE selected the most different set from the others. Although it also covered diet information and Mothers' exercise habit, the choices of the exact variables were different.

The model predictive performance was evaluated on the same testing set (Table 3) based on two measures, accuracy and weighed F-score. Ridge, Lasso and Filter method had good performance with both measures on the high level, SVM-RFE performed at a moderate level as the accuracy was satisfactory but the weighted F-score was only acceptable, and Random Forest had the lowest scores of both measures. Based on the performance and model property, the five models were classified into three levels with different weights (Table 4). Although the performance of the Filter method was relatively good, it failed to consider variable interactions, leading to declination in rank.

Table 3. Model predictive performance

	Filter (MI)	SVM-RFE	Ridge	Lasso	Random-Forest
Accuracy	0.843	0.845	0.844	0.839	0.828
Weighted F Score	0.915	0.774	0.915	0.912	0.770

Table 4. Model weights defined by performance

	Models	Model Weights
Level 1	Lasso regression, Ridge regression	1
Level 2	Filter (MI), SVM-RFE	0.5
Level 3	Random Forest	0.2

The voting scores for 3 voting strategies were calculated (Table 5) and the trend of the changes in feature importance can be observed more intuitively in Fig. 1, which shows the percentage of the score of one feature out of the whole set. One minor modification in the aggregated result was that PhysicalExercise_Adequate having a negative effect on the outcome was incorporated in PhysicalExercise_Inadequate. Some features gradually gained more importance during the changes from Voting1 to Voting3, including Age, Sex, Tobacco_No, DietEducation, and BreaskfeedingType_Maternal, which were the top 5 features selected by Voting3. MothersDiastolicPressure, on the opposite, lost its superiority. PhysicalExercise_Inadequate had a similar trend although it was still among the top 10 features of Voting3. MothersDietEducation was more stable and took almost the same share of voting scores in all strategies.

The top 10 features selected by BFSMR included age, sex, birth year, breastfeeding type, smoking habit and diet-related knowledge of both the children and their mothers, exercise, and Mother's systolic blood pressure. The results indicate that smoking habit, lack of exercise, and unbalanced diet of both mothers and children are the risk factors

of childhood obesity. Besides, boys have higher risk than girls and the risk grow along with age. It is also found that maternal breastfeeding can reduce the risk and younger generation tend to suffer more from obesity.

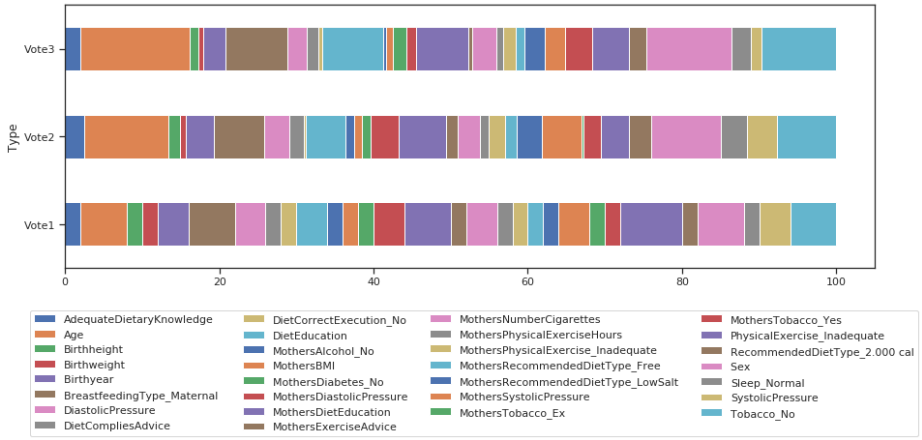


Fig. 1. Stacked bar plot of feature importance between 3 voting strategies shown in percentages

Table 5. Voting scores of different strategies.

Voting1		Voting2		Voting3	
PE_Inadequate	4	Age	30	Age	25
Age	3	Sex	25	Sex	19.4
BFType_Maternal	3	Tobacco_No	21	Tobacco_No	17
MoDietEducation	3	BFType_Maternal	18	BFType_Maternal	14
Sex	3	MoDietEducation	17	DietEducation	14
Tobacco_No	3	DietEducation	14	MoDietEducation	12
Birthyear	2	MoSystolicPressure	14	PE_Inadequate	8.4
DiastolicPressure	2	SystolicPressure	11	MoTobacco_Yes	6
DietEducation	2	Birthyear	10	MoNumberCigarett	5.5
MoDiastolicPressure	2	MoDiastolicPressure	10	Birthyear	5.2
MoNumberCigarettes	2	PE_Inadequate	10	MoSystolPressure	4.6
MoSystolicPressure	2	DiastolicPressure	9	DiastolicPressure	4.5
SystolicPressure	2	MoRDType_LowSalt	9	MoRDType_Low-Salt	4.5
AdeDKnowledge	1	Sleep_Normal	9	Sleep_Normal	4.5
Birthheight	1	MoNumberCigarettes	8	RDType_2.000 cal	4
Birthweight	1	RDType_2.000 cal	8	AdeDKnowledge	3.5
DietCompliesAdvice	1	AdeDKnowledge	7	MoDiabetes_No	3
DCExecution_No	1	MoPE_Inadequate	6	MoPE_Inadequate	3
MoAlcohol_No	1	MoTobacco_Yes	6	DietComplieAdvice	2.5
MoBMI	1	DietCompliesAdvice	5	SystolicPressure	2.5
MoDiabetes_No	1	Birthheight	4	MoDiastoPressure	2.3
MoExerciseAdvice	1	MoExerciseAdvice	4	Birthheight	2
MoPE_Inadequate	1	MoRDType_Free	4	MoRDTyp_Free	2

MoPEHours	1	MoAlcohol_No	3	MoBMI	1.5
MoRDType_LowSalt	1	MoBMI	3	MoPEHours	1.5
MoRDType_Free	1	MoDiabetes_No	3	Birthweight	1
MoTobacco_Yes	1	MoPEHours	3	DCExecution_No	1
MoTobacco_Ex	1	Birthweight	2	MoExerciseAdvic	0.8
RDType_2.000 cal	1	DCExecution_No	1	MoAlcohol_No	0.6
Sleep_Normal	1	MoTobacco_Ex	1	MoTobacco_Ex	0.2

Note: the same feature abbreviation as Table 2.

5 Discussion

5.1 Related Work

There are numerous studies on the risk factors for childhood obesity. In general, obesity is often considered to be the result of an imbalance between calories taken in and burned out. However, there is increasing evidence that other factors such as genetic background play a key role in determining the risk of obesity [24]. According to the review in 2017 [22], childhood obesity is the result of an interaction between different factors such as the environment, genetics and a child's surroundings. Environmental factors include lifestyle factors such as eating behaviors, which is highly related to parents' feeding styles, physical activity, stress and depression [10]. Other major environmental factors include perinatal factors [6, 23], birth size [31], catch-up growth [27], environmental chemicals [29], microbiota [20, 4] and adverse life experiences [11]. There are some studies that have demonstrated an association between the sleep time duration and obesity [19, 25].

A study in 2001 suggest that the main risk factors for obesity in children include dietary intake, physical activity and sedentary behaviour, moderated by factors such age and gender. In addition, family characteristics, parents' lifestyles and environmental factors (e.g. school policies and demographics) have a major impact in children's lifestyle and, therefore, their risk of obesity [7].

Another study in 2013 used multiple regression analyses to identify childhood obesity risk factors from data collected in a longitudinal study of preschool children and they conclude that the three early-life risk factors are parent BMI, child sleep duration and parental restrictive feeding [9]. Hammond's research predicted childhood obesity using electronic health records and publicly available data by means of a variety of machine learning algorithms [14].

5.2 Discussion

In this section, we discuss the advantages of the new method BFSMR and why it is useful in the real world setting.

The voting strategy played an important role in selecting a more reasonable feature set with better clinical relevance. The Bagging method uses the majority voting with equal probability like Voting1. But the scores of Voting1 were so close that it was difficult to distinguish the most important ones. For example, 13 features had scores higher

than 2, making it impossible to select only the top 10 features as the final output. Voting2 added feature weights based on the ranking from each model, nevertheless, variables with high rankings from poorly-fitted models could still affect the results. One significant difference between the scores of Voting2 and Voting3 was the notable decline in the rankings of the numeric variables of less relevance, such as MothersSystolic Pressure, SystolicPressure, MothersDiastolicPressure. The relevant numeric variables were not negatively affected by the voting strategy change, on the opposite, the ranking of MothersNumberCigarettes climbed from the 15th in Voting2 to the 9th in Voting3.

Furthermore, compared with other feature selection models, BFSMR selected variables that were easier for the follow-up interventions or policy decision-making process to implement, after analyzing each variable's importance through a better voting strategy. For example, although three models (SVM-RFE, Ridge, Lasso) all covered diet-related information, various variables were used in different models, e.g. RecommendedDietType_2000 cal, DietCompliesAdvice, DietCorrectExecution_ No. These variables were concrete about one specific aspect but it would be difficult to use them to make corresponding policies or interventions. The features selected by BFSMR were DietEducation and MothersDietEducation, which would be easier for practice in real life.

Another advantage of BFSMR is that it provides a general approach to deal with large-scale data and combine results from multiple models and it is flexible because the feature selection methods applied in this model are not defined. In our study, we chose 5 models as representatives of filter, wrapper, and embedded methods, but it is also possible to apply other feature selection methods when it is necessary.

Finally, although the results of single model were not reported as the final output, the process of applying various models was a comprehensive exploration of important features. Some variables only appeared once in one model, which might be neglected if only one model was applied. For instance, Sleep_Normal had a negative effect in the Filter method, which was in agreement with findings in other studies [19, 25]. Other interesting variables include MothersBMI and MothersDiabetes_No, indicating the effect of genetic factors. MothersRecommendedDietType_LowSalt and MothersRecommendedDietType_Free suggested some ideal diet styles to prevent childhood obesity. These variables were not included in the current output, but it would be very likely these variables were selected if we chose a higher number of selected features, for instance, Sleep_Normal and MothersRecommendedDietType_LowSalt ranked 12 in the voting scores and MothersDiabetes_No was at the 17th.

6 Conclusion

We presented BFSMR, a new machine learning approach that combines the framework of MapReduce and Bagging to achieve time efficiency on large-scale, high-dimensional data, and applies a collection of feature selection models to avoid model bias. The paper contributes with respect to existing work by being one of the first to comprehensively compare the model inclination to the selected features and combine multiple feature selection methods to give a model-averaging feature selection set instead of providing

predicted values or classification as the output. In addition, the paper addresses the interpretability and clinical relevance of the selected features rather than relying solely on statistical measures or predictive performance to select features.

We evaluated BFSMR on a dataset of 1,478,857 records and 119 features collected from 426,813 children in the Basque Country (Spain) from multiple data sources. BFSMR selects a set of features covering age, sex, birth year, breastfeeding type, smoking habit and diet-related knowledge of both the children and their mothers, exercise, and Mother's systolic blood pressure. The results suggest that a healthier life style of both mothers and their children, with more exercises, more balanced diet, and less cigarettes. Boys at higher age and younger generation suffer more from childhood obesity and maternal breastfeeding can reduce the risk. One limitation of this study is that the predictive performance of the five feature selection models used in the paper was not very satisfactory due to the sparse structure and missing values. Nonetheless, BFSMR acts as a general strategy to provide the framework of a meta-algorithm. The feature selection models can be replaced if particular data issues need to be solved.

References

1. Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E., Khalili, D.: A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. Journal of Clinical Epidemiology 71, 76 – 85 (2016).
2. Breiman, L.: Bagging Predictors. Machine learning 24, 123 – 140 (1996).
3. Breiman, L.: Random Forests. Machine Learning 45(1), 5-32 (2001).
4. Chang, L., Neu, J.: Early Factors Leading to Later Obesity: Interactions of the Microbiome, Epigenome, and Nutrition. Current Problems in Pediatric and Adolescent Health Care 45(5), 134-42 (2015).
5. Cover, T., Thomas, J.: Elements of Information Theory. Wiley; 1991.
6. Davis, E., Lazdam, M., Lewandowski, A., Worton, S., Kelly, B., Kenworthy, Y., Adwani, S., Wilkinson, A., McCormick, K., Sargent, I., Redman, C., Leeson, P.: Cardiovascular Risk Factors in Children and Young Adults Born to Preeclamptic Pregnancies: A Systematic Review. Pediatrics 129(6), e1552-e1561(2012).
7. Davison K., Birch L.: Childhood overweight: a contextual model and recommendations for future research. Obes Rev. 2(3), 159-71 (2001).
8. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: the Sixth Symposium on Operating System Design and Implementation 2004, Communications of the ACM,vol 51, 137-150. Association for Computing Machinery, San Francisco, CA (2004).
9. Dev D., McBride B., Fiese B., Jones B., Cho H.: Behalf Of The Strong Kids Research Team. Risk factors for overweight/obesity in preschool children: an ecological approach. Child Obes. 9(5), 399-408 (2013).
10. El-Behadli, A., Sharp, C., Hughes, S., Obasi, E., Nicklas, T.: Maternal depression, stress and feeding styles: Towards a framework for theory and research in child obesity. British Journal of Nutrition 113(S1), S55-S71 (2015).
11. Fuemmeler B., Dedert E., McClernon F., Beckham J.: Adverse childhood events are associated with obesity and disordered eating: results from a U.S. population-based survey of young adults. J Trauma Stress 22(4), 329-33 (2009).

12. Garavaglia, S., Sharma, A.: A Smart Guide to Dummy Variables: Four Applications and a Macro. In: Proceedings of the Northeast SAS Users Group Conference (1998).
13. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning 46(1-3), 389-422 (2002).
14. Hammond R, Athanasiadou R, Curado S, Aphinyanaphongs Y, Abrams C, Messito MJ, Gross R, Katzow M, Jay M, Razavian N, Elbel B.: Predicting childhood obesity using electronic health records and publicly available data. PLoS One 2019 14(4), e0215571 (2019).
15. Harris, D., Harris, S.: Digital design and computer architecture. San Francisco (CA): Morgan Kaufmann; 2012. p. 129
16. Hira, Z., Gillies, D.: A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Advances in Bioinformatics 2015, 1-13 (2015).
17. Hoerl, A., Kennard, R. Ridge regression: Biased estimation for nonorthogonal problems. Technomitrics 12, 55-67 (1970).
18. Jacobsen, B., Aars, N.: Changes in waist circumference and the prevalence of abdominal obesity during 1994–2008 - cross-sectional and longitudinal results from two surveys: the Tromsø Study. BMC Obes 3, 41 (2016).
19. Jiang, F., Zhu, S., Yan, C., Jin, X., Bandla, H., Shen, X.: Sleep and Obesity in Preschool Children. The Journal of pediatrics 154(6), 814-8 (2009).
20. Kalliomäki, M., Collado, M., Salminen, S., Isolauri, E.: Early differences in fecal microbiota composition in children may predict overweight. The American Journal of Clinical Nutrition 87(3), 534-538 (2008).
21. Kraskov, A., Stogbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E69, 066138. (2004).
22. Kumar, S., Kelly, A.: Review of Childhood Obesity. Mayo Clinic Proceedings 92(2), 251 – 265 (2017).
23. Lau E., Liu J., Archer E., McDonald S., Liu J.: Maternal weight gain in pregnancy and risk of obesity among offspring: a systematic review. J Obes. 2014:524939 (2014).
24. Sahoo K., Sahoo B., Choudhury A., Sofi N., Kumar R., Bhadoria A.: Childhood obesity: causes and consequences. J Family Med Prim Care 4(2), 187-92 (2015).
25. Sekine, M., Yamagami, T., Handa, K., Saito, T., Nanri, S., Kawaminami, K., Tokui, N., Yoshida, K., Kagamimori, S.: A dose–response relationship between short sleeping hours and childhood obesity: results of the Toyama Birth Cohort Study. Child: Care, Health and Development 28, 163-170 (2002).
26. Strobl, C., Boulesteix, A., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25 (2007).
27. Taveras E., Rifas-Shiman S., Sherry B., Oken, E., Haines, J., Kleinman, K., Rich-Edwards, J., Gillman, M.: Crossing Growth Percentiles in Infancy and Risk of Obesity in Childhood. Arch Pediatr Adolesc Med 165(11), 993–998 (2011).
28. Tibshirani, R.: Regression Shrinkage and Selection via the lasso. Journal of the Royal Statistical Society 58(1), 267-288 (1996).
29. Warner M., Wesselink A., Harley K., Bradman A., Kogut K., Eskenazi B.: Prenatal exposure to dichlorodiphenyltrichloroethane and obesity at 9 years of age in the CHAMACOS study cohort. Am J Epidemiol 179(11), 1312-22 (2014).
30. World Health Organization, Data and Statistics, http://www.euro.who.int/en/health-topics/noncommunicable-diseases/obesity/data-and-statistics, last accessed 2020/03/03 [22]
31. Yu, Z., Han, S., Zhu, G., Zhu, C., Wang, X., Cao, X., Guo, X.: Birth weight and subsequent risk of obesity: a systematic review and meta-analysis. Obesity Reviews 12, 525-542 (2011).