# Proceedings

## 5th Stochastic Modeling Techniques and Data Analysis International Conference with Demographics Workshop

# SMTDA2018

*Editor*

**Christos H Skiadas**

**12 - 15 June 2018**

**Cultural Centre of Chania, Crete, Greece**

# Calculation Methods for Binary Classification based on Discrete Data. An Application on Synthetic and Real Data.

Fragkiskos G. Bersimis[1], Iraklis Varlamis[1], Malvina Vamvakari[1]

and Demosthenes B. Panagiotakos[2]

[1] Department of Informatics and Telematics, Harokopio University, 9 Omirou Str., Athens, Greece;

[2]Department of Nutrition Science - Dietetics, Harokopio University, 70 Eleftheriou Venizelou Str., Athens, Greece

**ARTICLE HISTORY**

Compiled July 17, 2018

**ABSTRACT**

**Objective**: This work explores the performance of popular classification methods and multivariate indices on binary classification tasks, for datasets that comprise discrete valued features. This is directly applicable in the evaluation of the diagnostic accuracy of composite health related indices or screening tests, which combine multiple discrete-valued attributes (variables), usually using a weighted sum. **Methodology**: Several classification methods (e.g. logistic regression, classification trees, neural networks, support vector machines, ensemble classifiers etc) and multivariate indices that combine feature weighting techniques, are evaluated in this study using both simulated and actual medical-dietary data collected from the "*ATTICA study*" in Greece. A variety of scenarios that modify the discrete values' distribution parameters of the variables, and the number of variables as well as, are tested. All methods were assessed as to their classification performance by using a set of classification validity criteria such as: area under the ROC curve, true positive and true negative rates, positive and negative predictive value. The predictability of methods and the statistical significance of the results are evaluated using Monte-Carlo cross validation. **Results**: Results indicate that specific classification methods outperform all others in almost all the validity criteria and they also perform better than multivariate indices in certain cases, with regards to the data distribution, the number of features used and the number of their possible values. However, multivariate indices demonstrate a better performance when the number of features is small and the number of possible values in these features is also small.**Conclusion**: This work's findings propose a methodology for selecting more suitable techniques for predicting the clinical status of a person in the case of general or specific populations, depending on the data nature.

**KEYWORDS**

Binary Classification; Discrete Variables; Ordinal Features; ROC; AUC; Logistic Regression; Classification Trees; Neural Networks; Support Vector Machines, Classifier Ensemble

## 1. Introduction

The binary classification of living beings (e.g. to health or unhealthy), based on characteristics measured on a discrete scale, is an objective of many different scientific fields,

CONTACT F. G. Bersimis. Email: fbersim@hua.gr





such as medicine, psychometry, dietetics, etc (Carlsson, 1983; Jackson, 1970; Kant, 1996). This dichotomous classification was traditionally performed in health sciences with the aid of health indices (Bach et al., 2006; Beck et al., 1961; McDowell, 2006). Health related indices are quantitative variables that holistically assess a person's clinical condition by converting information usually from a variety of different attributes into a single-dimensional vector.

Discrete health indices are produced by the sum of discrete component variables that may be derived from discrete or continuous scale variables. An example of a discrete-scale variable is the number of cardiovascular events experienced by a patient and an example of a continuous-scale variable is the body mass index, which for convenience is appropriately categorized as "fat", "normal", "overweight" and "obese" with corresponding limits proposed by official health organizations, creating a hierarchical variable. Because of the ease of evaluating a feature in a discrete way, discrete scales are widely used (e.g., it is difficult for a person to accurately measure his training intensity per day, while it is easier to describe as mild, moderate or intense) although they provide less valid results than the continuous scales (Likert, 1952).

Although data mining is almost at the end of its third decade of research, and it has become popular in various fields during the last two decades, it is only recently that health scientists have invested on it (Tomar and Agarwal, 2013; Yoo et al., 2012). This is probably because supervised data mining techniques, such as classification and regression, need a lot of data to be trained and achieve a comparable performance to existing health indices, so they apply only on large cohort studies (Austin et al., 2013; Boucekine et al., 2013), or data from medical registries (Delen et al., 2005; Varlamis et al., 2017). It is also because of the limited interpretability of certain data mining based models, which in turns limits their applicability in certain cases. Classification and regression, are the two techniques that have been mostly applied on medical data in order to classify cases (Tang et al., 2005) or predict risks (Bottle et al., 2006), whereas clustering (Khanmohammadi et al., 2017; Yelipe et al., 2018) and association rules (Doddi, 2001; Sanida and Varlamis, 2017) are applied more rarely and mainly for their descriptive capabilities, that let researchers better understand or pre-process the dataset in hand.

The aim of the current study is to evaluate the performance of health related indices and classification algorithms under various dataset setups, given that they comprise only discrete valued features. For this, we comparatively examine classification methods and health related indices in terms of their classification accuracy on general population datasets, which comprising patients and non-patients. The research question is whether data mining (classification) methods can improve the sensitivity and specificity of existing health related indices (Kourlaba and Panagiotakos, 2009), in what extend and under which conditions. Synthetic and real data are used to study the aforementioned research question.

Since health indices are constructed specifically for each specific health case and dataset, in this work, we introduce a methodology for the data driven creation of composite indices. Their performance is compared against some well-known classification methods such as logistic regression, classification (decision) trees, random forests, artificial neural networks, support vector machine techniques and nearest neighbors classifiers as well as an ensemble classifier (meta-classifier) that combines all the previous methods.

In summary, the main contributions of this work are:

- a generic methodology for the construction of composite health indices for the

classification of datasets with discrete valued features,
- the evaluation of classification ensemble methods that combines more than one classifiers in order to improve individual classifiers performance,
- the evaluation of plain and ensemble classification methods and composite health related indices on synthetic and real datasets, with varying features.
- an open source software solution for the generation and evaluation of synthetic datasets that comprise discrete valued features, which can be used by future researchers to validate and extent the results of our study.

In section 2 that follows some related work is provided in order to identify similarities and differences with previous efforts in the recent literature. In section 3 the weighting process for the multivariate indices and the classification methods employed in the study are briefly described. Section 4 explains how the synthetic data were generated, how the "ATTICA study" data were collected and what evaluation criteria have been used in this study. Section 5 presents the results on synthetic data by providing the performance of the classification algorithms and indices for a varying number of discrete values and features, for a varying population size and ratio between diseased and healthy, as well as for different distribution parameters used. In section 6 the results of our work are discussed and an interpretation from a methodological perspective is attempted and section 7 summarizes our findings and concludes with directions for future work.

## 2. Related Work

Most of health related indices are combinations of individual attributes designed to measure specific medical and behavioral characteristics that are ambiguous or, in some cases, even impossible to be quantified directly and objectively (Bansal and Sullivan Pepe, 2013). There is a variety of clinical situations that cannot be measured with absolute precision, such as depression, anxiety, pain sensation of a patient, and the quality of eating habits (Huskisson, 1974; Trichopoulou et al., 2003; Zung, 1965). For clinical features such as the aforementioned there is a need of appropriate methods/tools to be discovered that quantify them on a discrete scale in order to classify individuals of a general population as patients or healthy. Even when the clinical features can be accurately measured with the appropriate measurement tools, such as hematological and biochemical markers, discretization contributes in the reduction of noise from the original readings (Ding and Peng, 2005).

Composite indices measure specific clinical features by using a suitable cut-off point (e.g., optimal separation point (Youden, 1950)). A health related index is usually synthesized by the sum of m component variables (features), where each of these features $X_i, i = 1, 2, ..., m$ expresses a particular aspect relative to the individual's clinical status. The scores of the m components are summed, with or without weighting, to provide an overall score. In the case of a composite health index $T_m$, the variables $X_i, i = 1, 2, ..., m$ can be either discrete or continuous. According to the index's value, the respective subjects examined are classified as either healthy or unhealthy, in terms of the appropriate diagnostic threshold for a particular disease (McDowell, 2006). In recent literature, several methods have been proposed to improve sensitivity, specificity and precision of these tools (Bersimis et al., 2013). More specifically, a health indice's diagnostic ability is improved by increasing the support of the component variables (Bersimis et al., 2017a) as well as by assigning weights to them (Bersimis et al., 2017b).

Composite health indices have been widely used in the medical field. For example, for predicting risk from cardiovascular disease by using mathematical/statistical models, explanatory variables such as age, gender, smoking, nutritional habits etc are associated with the existence of a chronic disease. Such indices have been used in prospective epidemiological studies (e.g. Framingham Heart) (Wang et al., 2003; Wilson et al., 1998), where the aggregation of the component variables provides the final index's score for the 10-year risk of a cardiovascular event (Dagostino et al., 2008). In the field of psychometry for the assessment of depression, there are a number of indices in the literature, such as the Hamilton Rating Scale for Depression (Hamilton, 1960) and the BDI (Beck Depression Inventory) (Beck et al., 1961). The aggregation of variable components provides the final index's score for depression estimation. The scoring of the above-mentioned indices is conducted by assigning high values in attitudes consistent with the condition of depression when they correspond to a high frequency and vice versa (Radloff, 1977). In the field of dietetics, a variety of indices have been constructed for evaluating the consumption's frequency and variety of food groups, such as the Diet Quality Index (DQI) (Patterson et al., 1994) and the Healthy Eating Index (HEI) (Kennedy et al., 1995).

For the classification of persons to patients and healthy, apart from the use of health indices that provide a univariate usually segregation approach, there are some well-known statistical multivariate methods. In particular, several statistical classification methods such as Logistic Regression (LR) (Vittinghoff et al., 2011), Classification and Regression Trees (CART) (Breiman et al., 1984), Neural Networks (NN) (Haykin, 1994)) and data mining elements such as machine learning and Support Vector Machines (SVM) (Kruppa et al., 2014) aiming at distinguishing two or more different groups in data sets that have a specific feature or not. The above methods have been developed mainly in the last decades when the application of Informatics' methods became an irreplaceable part of the medical research, resulting in the creation of Bioinformatics, which is a very wide interdisciplinary branch, aiming at studying and interpreting various biological phenomena. In addition, Biostatistics is the specialized scientific branch of Statistics that deals with the application of statistical methods, such as the management and analysis of numerical data, in the wider field of medicine and biological research.

Our work can be compared to (Maroco et al., 2011) since it extensively evaluates the performance of classification methods in terms of accuracy, sensitivity and specificity using a real dataset. In addition to this, we perform an extensive evaluation on synthetic data and provide the tool to future researchers for reproducing or extending our study. The current work extends previous works on the same dataset (the ATTICA study), which apply classification methods for risk prediction (Kastorini et al., 2013; Panaretos et al., 2018). However, in this study, it is the first time that an ensemble classification method that combines the merits of multiple classifiers is applied on the dataset.

## 3. Materials and methods

The main objective of this work is to evaluate the predictive performance of classification methods and health indices in the case of classifying a binary outcome variable based on discrete input variables. This section briefly presents the proposed health index construction methods, which apply to any dataset comprising discrete input variables and a binary output (Section 3.1), highlights the classification methods em-

ployed in our study (Section 3.2), and concludes with the proposed classifier ensemble method (Section 3.2.6), which considers all the available classifiers in tandem, in order to perform the binary prediction.

### *3.1. Data driven health index construction*

This study proposes a data driven composite indices' construction methodology, which targets on deriving the corresponding weighting formulas from logistic regression. More specifically, four discrete weighting methods $w_{ij}, i = 1, 2, ..., m, j = 1, 2, ..., 4$ for each component are proposed developed by using the odds ratios (OR) of univariate and multivariate logistic regression, as well as, by using the deviance statistic as modifying factor. These produced weighted indices $(T_1, T_2, T_3, T_4)$ are tested in simulated and real data. Moreover, weighted index $T_1$ is constructed by using the odds ratios of each component obtained from univariate logistic regression model $(OR_{ULR})$, whereas weighted index $T_2$ is constructed by using the odds ratios of each component obtained from multivariate logistic regression model $(OR_{MLR})$. Weighted indices $T_3$ and $T_4$ are constructed by using the aforementioned odds ratios in combination with the deviance statistic (DS) obtained from the corresponding logistic regressions. The deviance statistic (DS) i.e. the deviation between the theoretical model and the estimated model, is used for amplifying weights for the component variables that corresponds to lower deviation scores.Therefore, the weighted indices are defined by Equation 1:

$$T_j = \sum_{i=1}^{m} w_{ij} X_i, i = 1, 2, .., m, j = 1, 2, ..., 4 \tag{1}$$

where each $w_{ij}$ depending on the weighting method is given by the equations that follow:

$$w_{i1} = \frac{(OR_{ULR})_i}{\sum_{i=1}^{m} (OR_{ULR})_i}, w_{i2} = \frac{(OR_{MLR})_i}{\sum_{i=1}^{m} (OR_{MLR})_i}, \tag{2}$$

$$w_{i3} = \frac{(OR_{ULR}/DS)_i}{\sum_{i=1}^{m} (OR_{ULR}/DS)_i}, w_{i4} = \frac{(OR_{MLR}/DS)_i}{\sum_{i=1}^{m} (OR_{MLR}/DS)_i}, \tag{3}$$

$$\tag{4}$$

where $i = 1, 2, .., m$ corresponds to the components variables' multitude (Bersimis et al., 2017b).

### *3.2. Classification methods for discrete data*

#### *3.2.1. Logistic Regression*

The logistic regression model is a non-linear regression model applied in classification problems, where the dependent response variable $Y$ is categorical (not quantitative) with two or more categories. In the present study, a Binary Logistic Regression (where for example $Y$=1 means presence of a health risk and $Y$=0 means risk absence in a medical dataset) is applied. The simple logistic model is given by the following relation

(Vittinghoff et al., 2011):

$$P(Y = 1|X_j) = \frac{e^{\alpha+\beta X_j}}{1 + e^{\alpha+\beta X_j}} = \left(1 + e^{-(\alpha+\beta X_j)}\right)^{-1}, j = 1, 2, ..., m \quad (5)$$

where $P(Y = 1/X_j)$ express the conditional probability of a diseased individual.

### *3.2.2. Classification tree analysis*

Classification (decision) trees (Quinlan, 1986) constitute a highly interpretable machine learning technique, which uses a set of instances with known input and output variables to train a model, which can then be used to classify unknown instances. The learned models are represented using suitable graphs (tree form), which can also be interpreted as sets of rules (one rule for each path from root to the tree leafs) and can as well operate as decision models. As a prediction tool, classification trees are intended for problems that aim in predicting the right class for an unknown instance, choosing from one or more possible classes. During the training phase, they optimize the division of the known instances (training samples) to the tree leafs so that each leaf contains samples from the same class. The information gain (or KullbackLeibler divergence (Kullback and Leibler, 1951)) is one of the criteria employed to decide on the best split at each step. During the operation phase, the unknown instance is classified using the classification rules of the same tree and the label of the leaf defines its predicted class. Classification trees can work both with discrete and continuous data, although they usually discretize continuous feature in their pre-processing phase.

In this work, several input variables that correspond to discrete-valued dietary features are used in the real dataset and the output variable is also discrete and binary (the aim is to classify individuals as patients or not). However, when the input and output variables are continuous in nature, then it is possible to use regression tree analysis methods (e.g. CART (Breiman, 2017)) and learn the discretization limits of the output variable ((Bersimis et al., 2017b)).

Another limitation of decision tree methods is that they poorly operate in high-dimensional dataset, that comprise many features. Since the trees are usually shallow, they employ only a few of the features in their decision model with the risk to loose useful information from other features. For this reason, several multi-tree models, also known as $forests$, have been introduced in the literature and applied in classification problems, outperforming simple decision trees (see Random Forest (Liaw et al., 2002) and Rotation Forest (Rodriguez et al., 2006) algorithms). Such methods, are also known as classifier ensemble methods, since they combine more than one classifier in order to reach a decision. However, the classifiers in such ensembles are all of the same type (trees), whereas in this work, we experiment with a proposed mixed classified ensemble.

### *3.2.3. Bayesian (probabilistic) classifiers*

Probabilistic classifiers assume generative models, in form of product distributions over the original attribute space (as in naive Bayes) or more involved spaces (as in general Bayesian networks) (Kononenko, 2001). They output a probability for each unknown instance to belong to each of the classes and have been shown experimentally successful on real world applications (Pattekari and Parveen, 2012), despite the many simplified probabilistic assumptions. The Bayesian classifiers rely on Bayes' theorem,

which mainly assumes a strong (naive) independence between the input features.

Given an unknown instance to be classified, which is represented by a vector $x = (x_1, \ldots, x_n)$ in the space of $n$ features (independent variables), the classifier assigns to this instance probabilities: $p(C_k \mid x_1, \ldots, x_n)$ for each of $k$ possible outcomes or classes $C_k$.

For a large number of features ($n$), or for features with many discrete values the model based on probabilities is infeasible, since it will require too many instances to train (to learn probabilities). However, using Bayes' theorem, the conditional probability can be expressed proportionally to the product of all conditional probabilities of the classes given the feature values of the unknown instance.

$$p(C_k \mid x_1, \ldots, x_n) \propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{6}$$

As a result, the unknown instance $x$ is classified to the class $C_k$ that has the highest conditional probability according to equation 6.

### *3.2.4. Artificial neural networks*

Artificial neural networks (ANN) are applied in a variety of scientific fields such as medical diagnosis, speech & pattern recognition etc (Cho et al., 2014; Nigam and Graupe, 2004). ANN is a computing scheme representing partly the biological neural networks existing in human or animal brains, expressed by connected nodes (artificial neurons) organized properly in layers. All artificial neurons are connected and able to transmit signals, usually real numbers, through their connections (synapses) resulting to an output calculated suitably by a non-linear function according the initial inputs based on specific weights assigned to all neurons. ANN's greatest advantage is expressed by its ability to improve its performance by learning continuously by past procedures (Sutton et al., 1998).

### *3.2.5. Support Vector Machines*

Support vector machines are a supervised classification method, which is preferred for binary classification problems with high dimensionality (i.e. a large number of features) (Cortes and Vapnik, 1995). An SVM uses the training data , in order to build a model that correctly classifies instances with a non-probabilistic procedure. First, the space of the input samples, is mapped onto a high dimensional feature space so that the instances are better linearly separated. This transforms SVM learning into a quadratic optimization problem, which has one global solution. The optimal separating hyper plane in this new space must have the maximum possible margin from the training instances it separates from the two classes and the resulting formulation, instead of minimizing the training error seeks to minimize an upper bound of the generalization error. SVMs use non-linear kernel functions to overcome the curse of dimensionality (Azar and El-Said, 2014; Ding and Peng, 2005). They can handle both discrete and continuous variables as long as all are scaled or normalized. The ability of SVMs to handle datasets of large dimensionality (many features) made them very popular for medical data classification tasks. They are usually employed as is in binary classification tasks, but there is ongoing work on optimizations that can further improve SVM classifiers performance (Shen et al., 2016; Weng et al., 2016).

*3.2.6. Meta-classifier ensemble*

Ensemble classification methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions Dietterich (2000). Voting is the simplest form of a classifier ensemble. The main idea behind Voting is to use the majority vote or the average predicted probabilities given from conceptually different machine learning classifiers to predict the class labels. Such a classifier can be useful for a set of equally well performing model in order to balance out their individual weaknesses. Random and Rotation Forest algorithms are also considered ensemble methods, but they combine more classifiers of the same type (decision trees). Gradient Boosting (Friedman, 2001) is a meta-classifier that builds an additive model in a forward stage-wise fashion, which allows for the optimization of arbitrary differentiable loss functions. In each stage the algorithm trains a set of binary regression trees on the negative gradient of the binomial or multinomial deviance loss function. Gradient Tree Boosting (Hastie et al., 2001) or Gradient Boosted Regression Trees (GBRT) is a generalization of boosting to arbitrary differentiable loss functions. They have good predictive power and robustness to outliers in output space, but have increased complexity and phase scalability restrictions.

## 4. Experimental evaluation

This paragraph includes the methodology for the generation of synthetic data (section 4.1), the data collection method and the details of the ATTICA study dataset (section 4.2), as well as, the proposed methods accuracy evaluation measures (section 4.3). The code for generating the synthetic dataset and running the classification algorithms is available at BitBucket[1].

### *4.1. Synthetic data generation*

In order to evaluate the performance of composite indices and classifiers, we perform multiple tests, using various scenarios with regards to the input features, such as the distribution of each input variable and the number of their partitions, the number of samples in the population and the number of input variables in the dataset. For this reason, we developed a Python script for generating synthetic datasets, using several parameters, as explaining in the following.

First, we parametrized the variables' partitioning ($k$) (i.e. the possible values an input variable can take, ranging from 1 to $k$), which for simplicity was the same for all variables in our experiments[2].

Second, we employed a skewed discrete uniform distribution in all variables, with different mean ($meanpos, meanneg$) and deviation ($stdevpos, stdevneg$) for the distribution of diseased (positive) and non-diseased (negative) individuals. In our experiments we use the same shift of the mean (higher than the normal mean for positive samples and lower than the normal mean for negative samples), which however is proportional to $k$ ($meanshift = \frac{k+1}{2} + \lambda * (\frac{k+1}{2} - 1)$, where $\lambda$ defines the ratio of the shift). For example, the distribution of values (for positive and negative samples) in an attribute of the generated dataset for $k$=5, $stdevpos = stdevneg = 1$ and respectively

[1] `https://bitbucket.org/varlamis/discretedatagenerator`

[2] The code can be easily expanded to support the use of a vector or $k_i$ values, where $i$ is the number of input variables, instead of a single $k$

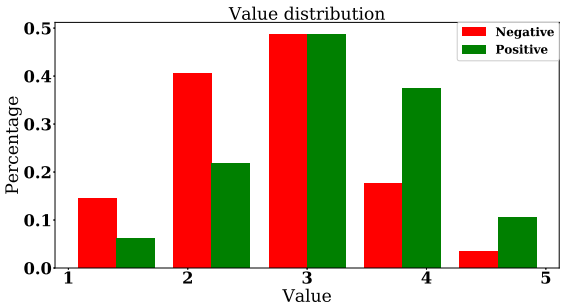


Figure 1.: Value distribution between positive and negative samples for $k$=5.

$mean_{pos} = 3.4$ and $mean_{neg} = 2.6$ is similar to that depicted in Figure 1.

Third, we variated the hypothetical population ratio between diseased and healthy individuals ($pos, neg$ respectively). Finally, we parametrized the population size ($samples$) and the number of input variables ($features$).

Modifying the aforementioned parameters leads to a dataset that simulates the dataset perspective of a real survey.

### *4.2. ATTICA study - Dietary data collection*

All methods and indices are also evaluated on real data, more specifically on data from the *ATTICA* epidemiologic study that took place in the Greek region of Attica within 2001 and 2002 (Pitsavos et al., 2003). At the beginning of the study, all participants were found healthy, free of any cardiovascular disease and during the study period, the consumption frequency of food groups was measured for the following food groups: cereals, fruits, nuts, vegetables, potatoes, legumes, eggs, fish, red meat, poultry, full fat dairy products, sweets and alcohol (measured in times/week consumed). From all participants in the ten-year follow up of the ATTICA study, we excluded those having missing values in any of the food groups, in order to avoid any missing values issues. From the 700 individuals that finally used in our study, 78 have reported a cardiovascular disease in the 10-years and 622 were categorized as healthy. This resulted in an unbalanced real dataset with the ratio of healthy to diseased being approximately 1:8.

The food consumption information was the only information used for classifying individuals to be healthy or non-healthy, with regards to the risk of occurrence of a cardiovascular disease within the 10-years period. More specifically, all variables corresponding to the aforementioned food groups were measured on a continuous scale, by counting portions per week. Then data were standardized using z-score and discretized by dividing the range of values into fixed-width intervals, depending on the desired number of partitions ($k$). This way, discrete data were produced with 3, 5, 7, 9 and 11 partitions each.

### *4.3. Evaluation of classification performance*

The diagnostic ability of a classification procedure is evaluated usually by using: i) accuracy (True Rate - TR) and the area (AUC) under the receiver operating characteristic curve (ROC), which is produced by mapping two-dimensionally the conditional probabilities Sensitivity (True Positive Rate - TPR) and 1- Specificity (True Negative

Rate - TNR), and ii) the Positive Predicted Value (PPV) and Negative Predicted value (NPV), in a specific cut off point. The value of Youdens J statistic (Youden, 1950) is a criterion for selecting the optimized cut off point of a diagnostic test, by maximizing the sum of sensitivity and specificity.

If we assume a random sample of diseased and non-diseased persons, who are classified by using a suitable discriminating method, four outcomes may occur that are presented in a 2x2 contingency table that includes:

- True characterized cases: the true positive cases (a) and the true negative cases (d).
- False characterized cases: the false positive cases (b) and the false negative cases (c).

Table 1.: 2x2 Contingency table for binary classification health cases. Green and red color indicates correct and incorrect classification, correspondingly.

| | | **True Clinical Status** | | |
|---|---|---|---|---|
| | | **Positive (Y=1)** | **Negative (Y=0)** | |
| **Predicted Clinical Status by the diagnostic test** | **Positive (Diseased)** | (a) True Positive Cases (TP) | (b) False Positive Cases (FP) | **a+b** |
| | **Negative (Healthy)** | (c) False Negative Cases (FN) | (d) True Negative Cases (TN) | **c+d** |
| | | **a+c** | **b+d** | **a+b+c+d=N** |

A test's sensitivity expresses the conditional probability of positives cases that are correctly identified as such, whereas specificity expresses the conditional probability of negative cases that are correctly identified as such. In addition, a test's positive predicted value expresses the conditional probability that a person with a positive examination is truly ill, and, negative predicted value expresses the conditional probability that a person with a negative examination is truly healthy (Daniel and Holcomb, 1995). Finally, accuracy expresses the conditional probability of positives or negative cases that are correctly identified as such (Daniel and Holcomb, 1995). The prediction accuracy was evaluated by using cross validation methods, such as 10-fold or Monte Carlo with a large number of repetitions and a randomized split (e.g. with 70:30 training/test ratio). More specifically, the prediction performance was evaluated for each method by separating initial synthetic data into training set and test set by using each partitioning technique. The process was performed 100 times and AUC average values are presented, along with their confidence intervals.
For evaluation purposes, we added two parameters that concern the train/test split ratio (*testpercentage*) and the number of repetitive (Monte Carlo) cross validations (*iterations*).

## 5. Results

### *5.1. Results on synthetic data*

The aim of the first experiment is to evaluate the performance of the different classification algorithms and multivariate indices, using several criteria. For this purpose, we use the dataset generator with specific parameters that simulate a typical case of a real world dataset with discrete valued attribute. We choose the number of possible

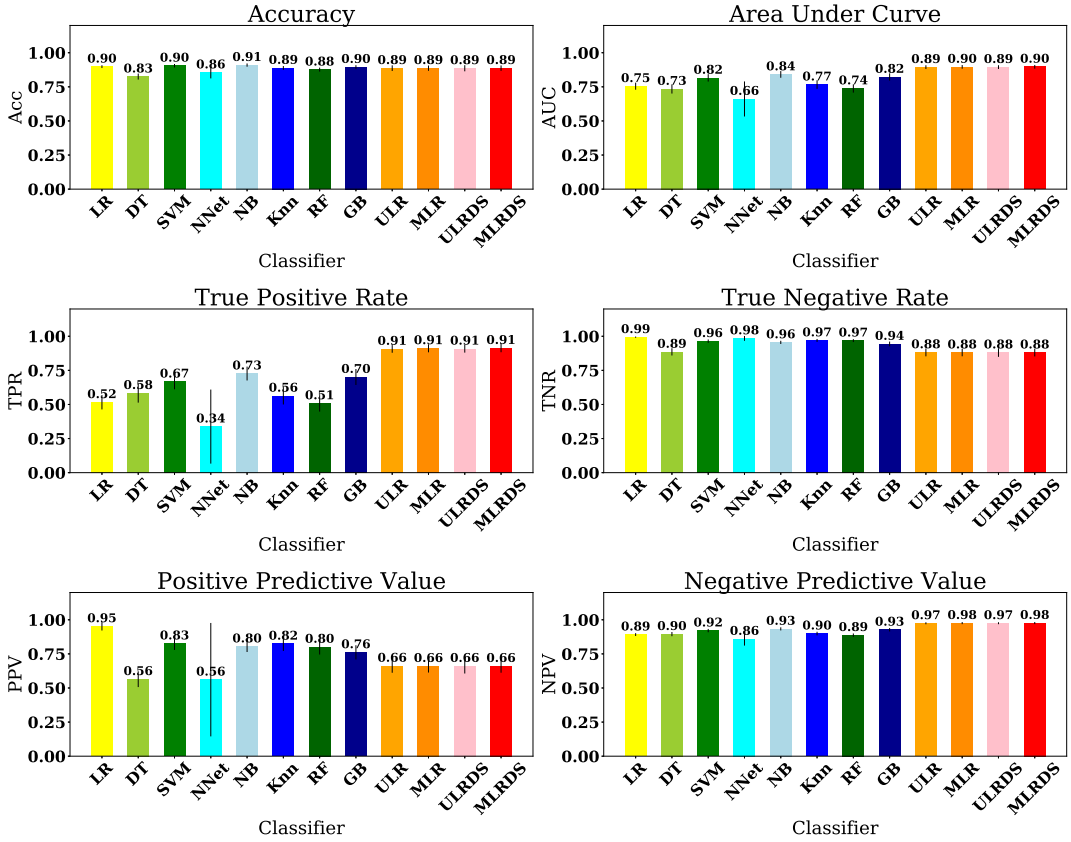


Figure 2.: The performance of the classification algorithms and indices.

values (we call them partitions) for all attributes to be from 1 to 5 (i.e. $k$=5), used a value of $\lambda = 0.2$ which results to $meanpos = 3.4$ and $meanneg = 2.6$ and the same standard deviation for positive and negative samples ($stdevpos = stdevneg$=1). The distribution of randomly generated values in the 10 features ($feat$=10) resembles that of Figure 1. We assumed a variety of samples with 1000 hypothetical individuals and a 1:4 positive to negative ratio (i.e. 200 patients and 800 healthy).

In the dataset that we generated, we repeated a random 70:30 train/test split 100 times and report the average values (and standard deviation). The results are summarized in the plots of Figure 2 that contain the six evaluation metrics (accuracy, AUC, sensitivity - TPR, true negative ratio - TNR, positive predictive value - PPV and negative predictive value - NPV) for each of the classification algorithms (logistic regression LR, Decision trees - DT, Support Vector Machines -SVM, Multi Layer Perceptron neural network -NN, Gaussian Naive Bayes classifier - NB, k-nearest neighbors classifier - Knn, Random Forests - RF, Gradient Boost classifier - GB) and the multivariate indices (ULR, MLR, ULRDS and MLRDS). The default parameters have been employed for all classifiers[3] in order to avoid biasing the results, with parameter tuning.

The results of Figure 2 show a good accuracy performance for all methods (0.81-0.89), with Naive Bayes (NB) having the highest accuracy from all methods and SVM

[3]We encourage reader to refer to the Sci-kit learn API documentation for more details on the default value parameters for each algorithm `http://scikit-learn.org/stable/modules/classes.html`

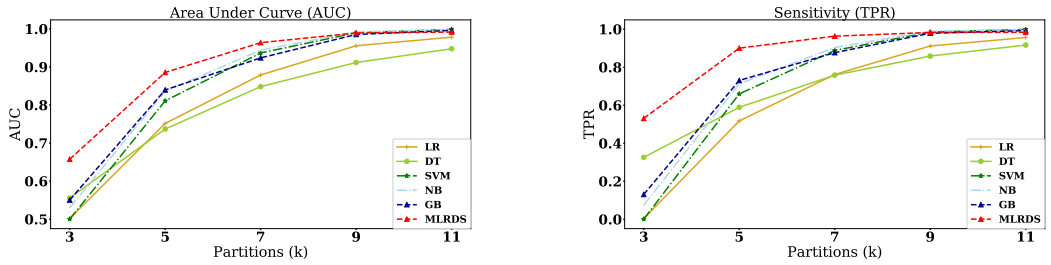


Figure 3.: AUC and Sensitivity for different $k$ values.

and Gradient Boost ensemble classifiers to follow. However, Naive Bayes suffers from low sensitivity, compared to other methods. On the contrary, the multivariate indices have a high sensitivity and high AUC values (the best among all methods), which is very important when searching for the minority of positives in a population. The multivariate indices suffer from low positive predictive values, which are probably due to the number of false positives they introduce.

In the second experiment, we keep all other values constant and modify the number of partitions ($k$ in the discretization step, or assuming that the discrete variables take values that range from 1 to $k$). Although we test all the algorithms, in Figure 3 we focus on the algorithms that performed better in the first experiment. From the results in Figure 3 it is obvious that as the number of partitions ($k$) increases, the AUC and sensitivity performance of the classifiers increase respectively. This was expected, since with more partitions (i.e. possible values for a discrete variable) the problem of class separation becomes easier. This finding is in agreement with earlier work by Bersimis (Bersimis et al., 2013) where it was proved that partition's increase corresponds to sensitivity increase. However, some algorithms always perform worse than others (e.g. decision trees and logistic regression perform worse than Gradient Boosting ensemble classifier, SVMs or Naive Bayes). Data driven indices such as $MLRDS$, which was constructed from the multivariate logistic regression model perform better than all other algorithms, even than Naive Bayes or SVM, although, for higher $k$ values there are no significant differences in their performance. In this particular set of experiments on synthetic data, the performance of the very simple method of Naive Bayes is extremely good. This happens because Naive Bayes is based on the naive assumption that the features are orthogonal (non correlated) to each other, which normally is not valid in a real dataset. The way we generated the synthetic dataset results in this orthogonality of features and justifies the high performance of Naive Bayes.

The third experiment examines the effect of the number of features ($feat$) in the classifiers' performance. For this purpose we repeat the experiments of datasets with 5, 10, 15, 20 and 50 features, using five discrete values ($k$=5) in all cases. From the results in Figures 4 we notice that decision trees cannot handle the high dimensionality of the dataset, which is a known restriction from the literature. Similarly, logistic regression demonstrates a low performance, which improves, but slightly, when the number of features increases. Ensemble classifiers such as Random Forests (not in the plots) and Gradient Boost manage to cover the high dimensionality by training more than one models with a subset of the dimensions each time, but still perform worse than SVMs. Finally, the performance of Naive Bayes (with the assumption of orthogonality) and SVM improves in high dimensions and outperforms that of multivariate regression indices. The latter are ideal for datasets with a few discrete valued features but reach

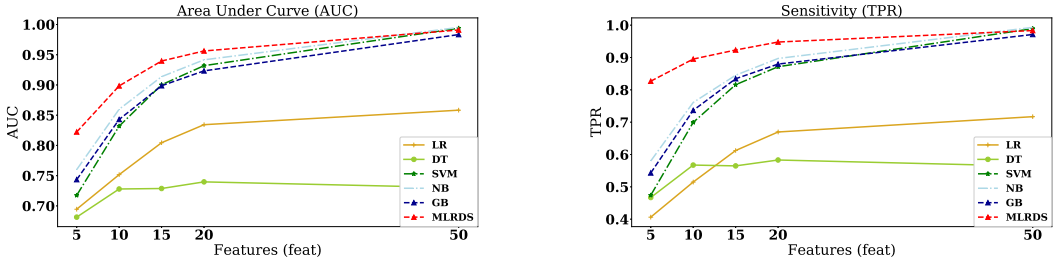


Figure 4.: AUC and Sensitivity for a varying features number.

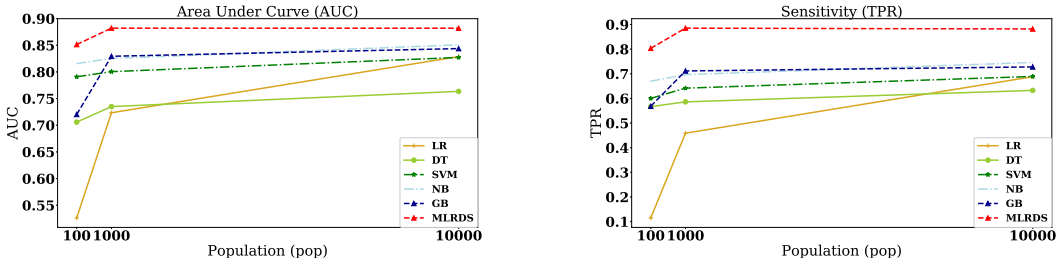


Figure 5.: AUC and Sensitivity for a varying population size.

a performance upper bound above 20 features.
The aim of the fourth experiment was to examine the effect of the population size to the performance of the different classification methods. Using the same configuration as in the first experiment but with a population varying from 100 to 10,000 instances we get the results depicted in Figure 5. The results show a significantly better performance for the MLRDS classifier, but all classifiers tend to improve their performance as the population size increases. The logistic regression method improves the most by this increase in the population size, which probably means that it needs more data to be trained than other methods. However, it is far from the performance of MLRDS.
The fifth experiment examines the effect of the ratio between healthy and patient samples in the dataset. It is very unusual in medical datasets to have a balance in the number of patient and healthy instances, and this adds restrictions to several classification methods. In this experiment, we keep the same configuration as in the first experiment but we modify the ratio of patient:healthy in the following values: $1:1$ (balanced), $1:2$, $1:4$, $1:9$. The results in Figure 6 show a drop in the performance of

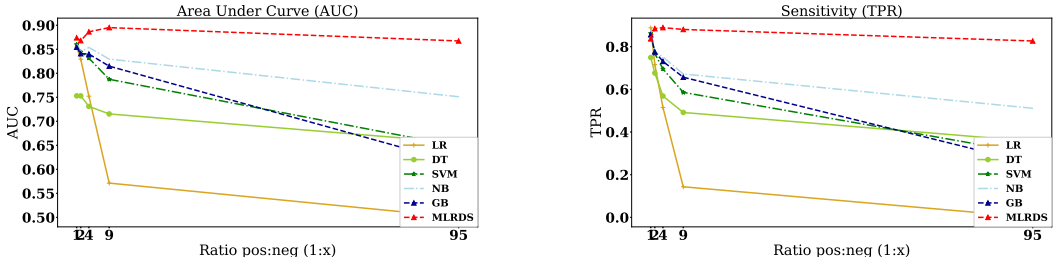


Figure 6.: AUC and Sensitivity for a varying positive:negative ratio.

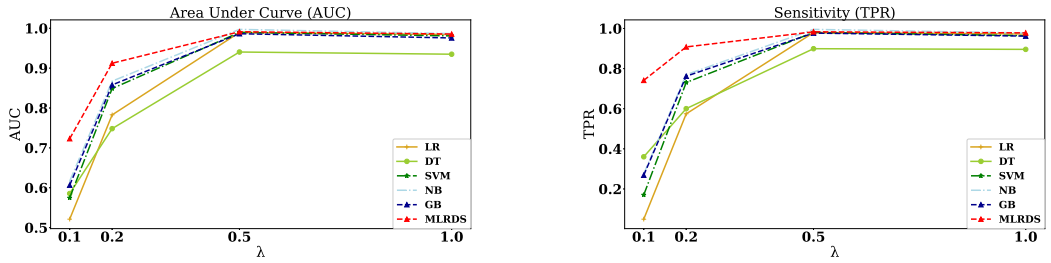


Figure 7.: AUC and Sensitivity for varying $\lambda$ values.

all algorithms for ratios lower than $1:4$ (25% patients in the dataset). It is interesting to note the increase in performance of MLRDS for the $1:4$ ratio (10% patients in the dataset) and its significantly better performance for highly unbalanced datasets (ratios $1:9$ or $1:95$).
The last experiment on synthetic datasets examines the effect of the separation of the distribution of feature values between positive and negative instances as determined by the $\lambda$ parameter. Once again, we keep the same configuration as in the first experiment, but modify $\lambda$ from 0.1 to 1. The results in Figure 7 show the poor performance of Logistic regression and Decision Trees for small $\lambda$ values, where the separation problem is harder. They also show that Multivariate indices achieve the best performance. We expect the classifiers' performance to improve, since the problem is easier when the distributions of values for positive and negative samples are well separated, and this happens for all methods. However, results show that Decision Trees perform worse than other methods for higher $\lambda$ values. This bad performance is probably due to the use of default parameters for the decision tree algorithm and can be possibly improved with the proper parameter tuning, which however is outside the scope of this work.

### *5.2. Results on real data*

The results of the evaluation of all algorithms on the ATTICA study data, are depicted in Figure 8.
Although the accuracy of data mining algorithms is higher than that of the multivariate indices, this is mainly due to their high true negative ratio (TNR). The performance of multivariate indices is more stable in all metrics and they demonstrate slightly higher AUC values than the data mining algorithms. More specifically, data mining algorithms seems to fail in the criterion of true positive ratio (TPR), whereas, achieve slightly greater values in negative predicted value (NPR). A more careful examination of the TPR subplot shows that Decision Tree classifier and Naive Bayes, which usually work better with discrete data, outperform all other data mining techniques in TPR and rank after the multivariate indices techniques in AUC.
Further experiments with less and more fine-grain discretization ($k$ from 5 to 51) shows that the sensitivity (TPR) of the multivariate indices[4] is in average much higher than that of the other classifiers, even of the Decision Trees classifier, that comes second. In terms of AUC, the MLRDS index and the Gradient Boost classifier ensemble are slightly better, but not significantly, than other methods, and have small fluctuations (in the third decimal) for higher $k$ values.

[4]only the MLRDS index is depicted in Figure 9, but all other indices behave similarly.

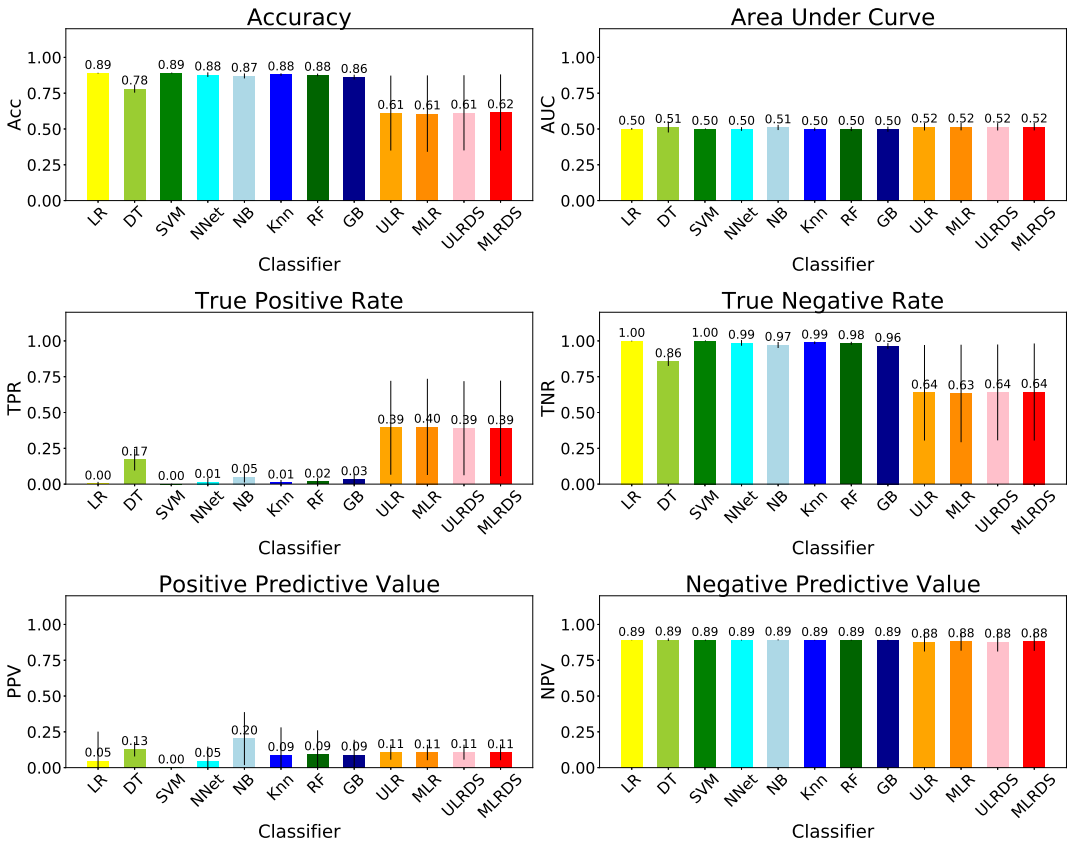


Figure 8.: The performance of the classification algorithms and indices on the data of the ATTICA study (k=7).

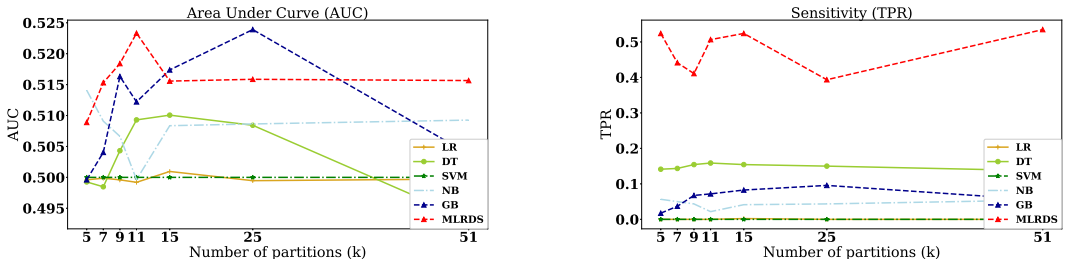


Figure 9.: AUC and Sensitivity for different discretization levels ($k$).

## 6. Discussion

Health indices are extensively used in health research fields such as cardiovascular risk prediction (Wang et al., 2003), depression evaluation (Zung, 1965) and nutritional assessment (Patterson et al., 1994) by measuring diseases' specific aspects and calculating a total score for classifying an individual as high or low risk etc. Classification methods are also used lately in health fields such as cardiovascular and cancer risk prediction by using data mining techniques Varlamis et al. (2017); Weng et al. (2016). Both tools aim to evaluate, classify and predict health conditions aiming to assist medical community to understand and interpret the mechanisms of various diseases.
In this work, classification methods and composite indices are compared in order to construct a methodological framework, which could assist any researcher aiming to conduct a classification procedure in medical data. The simulations' results by various scenarios performed in this work, showed differences in evaluation criteria for classification methods and indices used. More specifically, Naive Bayes (NB) classifier achieved the greatest value in accuracy, whereas the greatest value of the area under the receiver characteristic curve (AUC) and true positive ratio was achieved by the weighted indices. This shows an efficient performance of weighted indices in classification problems, when applied on data with similar characteristics as the simulated data of our study (equal value distribution and scale for all features, zero correlation between features, and imbalance between the two classes. The greatest value of true negative ratio was achieved by logistic regression and neural networks, therefore these methods could be conducted in special populations where high specificity is needed. The greatest value of positive predictive value was achieved by logistic regression and support vector machine. In contrary, the greatest value of negative predictive value was achieved by the weighted indices.
The simulations results revealed a significant increase in AUC and sensitivity of classification methods and weighted indices when the number of partitions increase above 7. For small values of k ($k < 7$), weighted indices seems to outperform classification methods, whereas for great values of k ($k > 7$) classification methods like SVM achieve greater scores in criteria AUC and sensitivity. This shows that classification methods can better handle features with many discrete values, which resemble to continuous features. Even among the classification methods, there exist many differences. For example, decision trees and logistic regression perform worse than Gradient Boosting Ensemble classifier, SVMs or Naive Bayes.
In addition, increase of the features' multitude led to the AUC increase except the method of decision trees in which the increase in the components seems to confuse the discretion of this method, which is noted in the literature ((Zekić-Sušac et al., 2014)). The results of our study in low and high dimensional spaces are in agreement with the related literature: i) we observe that for a small number of features, the weighted indices perform better than the classification methods, whereas when the number of features significantly increases, support vector machines (SVM) and Naive Bayes (NB) performed better (Bolivar-Cime and Marron, 2013), and ii) in all cases the increase rate is smaller for a larger number of features (Bersimis et al., 2017b).
Increasing the size of the population leads to an increase in the values achieved by many classification methods in the AUC and sensitivity evaluation criteria and, at the same time, to a reduction in the values achieved by the weighted indices in the same criteria. The highest increase rate is recorded by logistic regression, while weighted indices achieve higher values than classification methods, at any sample size. Therefore, the increase in the size of the population seems to have a more pronounced impact on

some classification methods.
A lower ratio of patients to healthy individuals results in a drop in performance for all methods. Thus, for highly unbalanced sets the classifiers performance is worst, i.e. the rarer a disease is, the more difficult it is to detect it. Indices show greater diagnostic ability in cases of very rare diseases, such as 1:95, and they also showed an increase in 1:4 case, in contrast to other methods.
When the distance between the theoretical *population means* of health and diseased individuals is relatively small, i.e. small $\lambda$ values, then the diagnostic ability of the weighted indices is low, but higher than the one of the classification methods, measured by AUC and sensitivity criteria. When the separation between diseased and non-diseased becomes easier, i.e. higher $\lambda$ values, the classification methods outperform weighted indices.
The better performance of composite indices in some of the setups in the artificial (synthetic) data is validated with the real data of ATTICA study. For example the composite indices have a larger AUC area and sensitivity than classification methods. However, the overall accuracy of classification methods is higher, and this is mainly because classification methods tend to produce more negatives (i.e. their Negative Predictive Value is close to 1).
The combination of a variety of medical (clinical, biological or behavioral) features, measured on a discrete scale, for classifying individuals of a general population as diseased or not, is an important process for establishing effective prevention strategies in various health areas, such as cardiovascular and cancer risk, metabolic disorders, malnutrition, risk of infant mortality, etc.
Conclusively, this work propose methods for the selection of an effective diagnostic method by using suitable classification methods or weighted indices in relation to the health data nature such as derived from psychological diseases or nutritional adequacy, etc. (McCullough et al., 2000). Moreover, the use of classification methods or weighted indices should be suggested for diagnostic procedures due to the fact that sensitivity and/or specificity increase in many cases shown as it is shown in previous paragraphs. In addition, further research is needed in this area because the classification method's accuracy and weighted indices' diagnostic ability have not been adequately studied.

## 7. Conclusions

Composite indices derived from multivariate methods seem to be sufficient solutions for classifying individuals in the case of discrete features with small partition number, since they perform better than classification algorithms. However, the latter are better for higher numbers of partitions k. Classification methods such as SVMs are preferable for high dimensional spaces i.e. for datasets with a big number of features/variables. In addition, in the case of orthogonal feature spaces, i.e. non-correlated variables, Naive Bayes classifier is a fast alternative that outperforms all other methods. In the case of available large training datasets, logistic regression is a well performing and fast alternative that competes other methods in evaluation criteria. For highly imbalanced datasets it is preferable to use multivariate indices than simple regressions methods or SVMs. Ensemble methods are also a good solution, since they combine multiple classifiers. Finally, for high $\lambda$ values, i.e. easy separable problems, SVMs and ensemble methods perform better than multivariate indices.
The next steps of our work in this field are to experiment with more real datasets, especially datasets that are inherently discrete. Also we plan to extend our synthetic

dataset generator to allow for different scales for each feature, different distributions and combinations of discrete and continuous features. Thus we will better simulate real datasets and will allow researchers to experiment with synthetic data of any size that resemble their real data. Finally, we will add more classification methods and optimization strategies, e.g. feature selection and parameter tuning in order to compile a powerful experimentation platform.

## 8. Acknowledgements

* The ATTICA study was supported by research grants from the Hellenic Cardiological Society (HCS2002) and the Hellenic Atherosclerosis Society (HAS2003).

## References

P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology*, 66(4): 398–407, 2013.

A. T. Azar and S. A. El-Said. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. *Neural Computing and Applications*, 24(5): 1163–1177, 2014.

A. Bach, L. Serra-Majem, J. L. Carrasco, B. Roman, J. Ngo, I. Bertomeu, and B. Obrador. The use of indexes evaluating the adherence to the mediterranean diet in epidemiological studies: a review. *Public health nutrition*, 9(1a):132–146, 2006.

A. Bansal and M. Sullivan Pepe. When does combining markers improve classification performance and what are implications for practice? *Statistics in medicine*, 32(11):1877–1892, 2013.

A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571, 1961.

F. Bersimis, D. Panagiotakos, and M. Vamvakari. Sensitivity of health related indices is a non-decreasing function of their partitions. *Journal of Statistics Applications & Probability*, 2(3):183, 2013.

F. Bersimis, D. Panagiotakos, and M. Vamvakari. Investigating the sensitivity function's monotony of a health-related index. *Journal of Applied Statistics*, 44(9):1680–1706, 2017a.

F. G. Bersimis, D. Panagiotakos, and M. Vamvakari. The use of components weights improves the diagnostic accuracy of a health-related index. *Communications in Statistics-Theory and Methods*, pages 1–24, 2017b.

A. Bolivar-Cime and J. Marron. Comparison of binary discrimination methods for high dimension low sample size data. *Journal of Multivariate Analysis*, 115:108–121, 2013.

A. Bottle, P. Aylin, and A. Majeed. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *Journal of the Royal Society of Medicine*, 99(8): 406–414, 2006.

M. Boucekine, A. Loundou, K. Baumstarck, P. Minaya-Flores, J. Pelletier, B. Ghattas, and P. Auquier. Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: a cohort study. *BMC medical research methodology*, 13(1):20, 2013.

L. Breiman. *Classification and regression trees*. Routledge, 2017.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and regression trees, chapman & hall. *New York, NY, USA*, 1984.

A. M. Carlsson. Assessment of chronic pain. i. aspects of the reliability and validity of the visual analogue scale. *Pain*, 16(1):87–101, 1983.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

W. W. Daniel and J. J. Holcomb. Biostatistics: a foundation for analysis in the health sciences. 1995.

D. Delen, G. Walker, and A. Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127, 2005.

T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

S. R. D. C. T. S. Doddi, Achla Marathe. Discovery of association rules in medical data. *Medical informatics and the Internet in medicine*, 26(1):25–33, 2001.

R. B. Dagostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro, and W. B. Kannel. General cardiovascular risk profile for use in primary care: the framingham heart study. *Circulation*, 117(6):743–753, 2008.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

M. Hamilton. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56, 1960.

T. Hastie, J. Friedman, and R. Tibshirani. Boosting and additive trees. In *The Elements of Statistical Learning*, pages 299–345. Springer, 2001.

S. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

E. Huskisson. Measurement of pain. *The lancet*, 304(7889):1127–1131, 1974.

D. N. Jackson. A sequential system for personality scale development. In *Current topics in clinical and community psychology*, volume 2, pages 61–96. Elsevier, 1970.

A. K. Kant. Indexes of overall diet quality: a review. *Journal of the American Dietetic Association*, 96(8):785–791, 1996.

C.-M. Kastorini, G. Papadakis, H. J. Milionis, K. Kalantzi, P.-E. Puddu, V. Nikolaou, K. N. Vemmos, J. A. Goudevenos, and D. B. Panagiotakos. Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-art classification algorithms: a case/case-control study. *Artificial intelligence in medicine*, 59(3):175–183, 2013.

E. Kennedy, J. Ohls, S. Carlson, and K. Fleming. The healthy eating index: design and applications. *Journal of the American Dietetic Association*, 95(10):1103–1108, 1995.

S. Khanmohammadi, N. Adibeig, and S. Shanehbandy. An improved overlapping k-means clustering method for medical applications. *Expert Systems with Applications*, 67:12–18, 2017.

I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

G. Kourlaba and D. Panagiotakos. The number of index components affects the diagnostic accuracy of a diet quality index: the role of intracorrelation and intercorrelation structure of the components. *Annals of epidemiology*, 19(10):692–700, 2009.

J. Kruppa, Y. Liu, G. Biau, M. Kohler, I. R. König, J. D. Malley, and A. Ziegler. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal*, 56(4):534–563, 2014.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

R. Likert. A technique for the development of attitude scales. *Educational and psychological measurement*, 12:313–315, 1952.

J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. de Mendonça. Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity

and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC research notes*, 4(1):299, 2011.

M. L. McCullough, D. Feskanich, E. B. Rimm, E. L. Giovannucci, A. Ascherio, J. N. Variyam, D. Spiegelman, M. J. Stampfer, and W. C. Willett. Adherence to the dietary guidelines for americans and risk of major chronic disease in men–. *The American journal of clinical nutrition*, 72(5):1223–1231, 2000.

I. McDowell. *Measuring health: a guide to rating scales and questionnaires*. Oxford University Press, USA, 2006.

V. P. Nigam and D. Graupe. A neural-network-based detection of epilepsy. *Neurological Research*, 26(1):55–60, 2004.

D. Panaretos, E. Koloverou, A. C. Dimopoulos, G.-M. Kouli, M. Vamvakari, G. Tzavelas, C. Pitsavos, and D. B. Panagiotakos. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002–2012): the attica study. *British Journal of Nutrition*, pages 1–9, 2018.

S. A. Pattekari and A. Parveen. Prediction system for heart disease using naïve bayes. *International Journal of Advanced Computer and Mathematical Sciences*, 3(3):290–294, 2012.

R. E. Patterson, P. S. Haines, and B. M. Popkin. Diet quality index: capturing a multidimensional behavior. *Journal of the American Dietetic Association*, 94(1):57–64, 1994.

C. Pitsavos, D. B. Panagiotakos, C. Chrysohoou, and C. Stefanadis. Epidemiology of cardiovascular risk factors in greece: aims, design and baseline characteristics of the attica study. *BMC public health*, 3(1):32, 2003.

J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

L. S. Radloff. The ces-d scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, 1(3):385–401, 1977.

J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1619–1630, 2006.

T. Sanida and I. Varlamis. Application of affinity analysis techniques on diagnosis and prescription data. In *Computer-Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium on*, pages 403–408. IEEE, 2017.

L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, and D. Liu. Evolving support vector machines using fruit fly optimization for medical data classification. *Knowledge-Based Systems*, 96:61–75, 2016.

R. S. Sutton, A. G. Barto, et al. *Reinforcement learning: An introduction*. MIT press, 1998.

T.-I. Tang, G. Zheng, Y. Huang, G. Shu, and P. Wang. A comparative study of medical data classification methods based on decision tree and system reconstruction analysis. *Industrial Engineering and Management Systems*, 4(1):102–108, 2005.

D. Tomar and S. Agarwal. A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5):241–266, 2013.

A. Trichopoulou, T. Costacou, C. Bamia, and D. Trichopoulos. Adherence to a mediterranean diet and survival in a greek population. *New England Journal of Medicine*, 348(26):2599–2608, 2003.

I. Varlamis, I. Apostolakis, D. Sifaki-Pistolla, N. Dey, V. Georgoulias, and C. Lionis. Application of data mining techniques and data analysis methods to measure cancer morbidity and mortality data in a regional cancer registry: The case of the island of crete, greece. *Computer methods and programs in biomedicine*, 145:73–83, 2017.

E. Vittinghoff, D. V. Glidden, S. C. Shiboski, and C. E. McCulloch. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer Science & Business Media, 2011.

T. J. Wang, J. M. Massaro, D. Levy, R. S. Vasan, P. A. Wolf, R. B. D'agostino, M. G. Larson, W. B. Kannel, and E. J. Benjamin. A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the framingham heart study. *Jama*, 290 (8):1049–1056, 2003.

Y. Weng, C. Wu, Q. Jiang, W. Guo, and C. Wang. Application of support vector machines in medical data. In *Cloud Computing and Intelligence Systems (CCIS), 2016 4th International Conference on*, pages 200–204. IEEE, 2016.

P. W. Wilson, R. B. DAgostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

U. Yelipe, S. Porika, and M. Golla. An efficient approach for imputation and classification of medical data values using class-based clustering of medical records. *Computers & Electrical Engineering*, 66:487–504, 2018.

I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36(4):2431–2448, 2012.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

M. Zekić-Sušac, S. Pfeifer, and N. Šarlija. A comparison of machine learning methods in a high-dimensional classification problem. *Business Systems Research Journal*, 5(3):82–96, 2014.

W. W. Zung. A self-rating depression scale. *Archives of general psychiatry*, 12(1):63–70, 1965.