Predictive Data Mining in Nutrition Therapy

Diana Ferreira
Informatics Department
University of Minho
Braga, Portugal
a72226@alunos.uminho.pt

Hugo Peixoto, José Machado and António Abelha
Algoritmi Research Center
University of Minho
Braga, Portugal
{hpeixoto,jmac,abelha}@di.uminho.pt

***Abstract*— The assessment and measurement of health status in communities throughout the world is a massive information technology challenge. Data mining, plays a vital role in health care industry since it really has the potential to generate a knowledge-rich environment that reduces medical errors, decreases costs by increasing efficiency, improves the quality of clinical decisions and significantly enhances patient's outcomes and quality of life. This study falls within the context of nutrition evaluation and its main goal is to apply classification algorithms in order to predict if a patient needs to be followed by a nutrition specialist. One of the tools resorted in this study was the Waikato Environment for Knowledge Analysis (Weka in advance) Workbench since it allows to quickly try out and compare different machine learning solutions. The tasks involved in the development of this project included data preparation, data preprocessing, data transformation and cleaning, application of several classifiers and its respective evaluation through performance measures that include the confusion matrix, accuracy, error rate, and others. The accomplished results showed to be quite optimistic presenting promising values of performance measures, specifically an accuracy around 91%.**

Keywords—information technology; data mining; health care; clinical decisions; nutrition evaluation; machine learning; classification algorithms; performance measures

I. INTRODUCTION

The past few years have witnessed a growing recognition of intelligent techniques like machine learning and data mining. Machine learning's purpose is the construction of computer systems that have the ability to adapt, learn and improve their performance in a given domain through experience [1][2]. Data mining is a multidisciplinary field that applies machine learning techniques, mathematical functions, and statistical analysis to discover interesting patterns or rules and extract previously unknown and potentially useful knowledge [3][4]. Data mining techniques include descriptive algorithms, for finding interesting patterns in the data, like associations, clusters, and subgroups [5], and predictive algorithms, that perform induction to make predictions of a specific attribute, which results in models that can be used for regression and classification [6][7]. Classification models predict categorical labels (discrete) while regression models predict continuous-valued functions [8]. Although data mining and knowledge discovery in databases (KDD in advance) are often treated as synonyms, in reality, data mining is only an important step in the KDD process. KDD is a field encompassing theories, methods, and techniques that maps low-level data which is often too voluminous to be understood into high-level knowledge that is more compact and useful [1][3]. The KDD process uses specific data-mining methods and machine learning algorithms, for pattern discovery and extraction [1][3]. This is an iterative multistep process that consists of selection, preprocessing, transformation, data mining, interpretation and evaluation, and knowledge representation [1][6].

A major challenge faced by healthcare organizations is the delivery of quality services at affordable costs. Quality service implies correct patient diagnosis and effective treatment delivery. However, clinical decisions are usually made based on doctors' intuition and experience rather than on the rich and potentially lifesaving knowledge stored in the healthcare databases, which results in undesired errors, biases, and excessive medical costs. Additionally, healthcare organizations store and collect large amounts of data on a daily basis, leading to more difficulties in analyzing data and slowing down the decision-making process. In critical environments, decisions need to be performed quickly. Thus, the automation of medical analysis and the development of a solution able to predict events before their occurrence would be highly beneficial for both patient and institution [9] [10] [11]. Data mining holds great promises for healthcare, since through its predictive features, it's possible to anticipate disease occurrence, progression, and prognosis [12][13], as well as, to improve physician's performance, enhancing the rational use of resources and consequently optimizing health care, allowing to identify high-risk patients and intervene proactively [5][10]. In hospital environment, when a physician suspects that a patient has an abnormal nutritional state, he sends a request to the nutrition service so that the nutritionist can analyze the patient status and decide whether or not he should have nutritional monitoring [14]. If the process time between the request and the answer from the nutritionist is substantial, the malnourished patient may not have the needed follow-up in time. Identifying the risk of malnutrition in patients from predictive variables is the first step towards an adequate nutritional control [14]. Given its prevalence, the traceability and monitoring of nutritional status should be available in the hospital environment to prevent, treat and improve its prognosis. With this, morbidity, mortality, as well as hospitalization time and hospital costs will be reduced, enhancing life quality. Given this reality, the nutritionist plays a crucial role, since he can identify early cases of nutritional risk [14]. In this way, the nutritionist can interfere with the control of patient's clinical status and, consequently, prevent and control its malnutrition, as well as infer the improvement of its clinical state. Thus, the application of data mining and machine learning techniques is

vital to help the nutritionist in the decision-making process. In the data mining process, typically one of the attributes is taken as the dependent attribute, representing the concept to be predicted by a pattern or a rule [3]. In this case, it is intended to predict the nutritionist's response to a request for patient follow-up by the nutrition service. This aims to reduce the process time between the request and the answer and, therefore, to ease the assessment and measurement of the nutritional health status providing an immediate and adequate treatment to the patient.

The dataset used in this project consists of clinical records obtained from a Portuguese hospital and contains information referring to evaluation records of nutrition episodes for patients with suspicion of nutritional imbalance, recording a period between 1st August, 2011 to 4th January, 2017. According to everything previously mentioned, the main goal of this study is to apply different data mining techniques in Weka environment to extract useful information from data and identify a suitable algorithm for generating an accurate predictive model to predict the need for patient follow-up by the nutrition service.

II. Practical Development Methodology

A. Data caracterization

The data used in the data mining process came from a data warehouse developed in a previous project [14]. This data warehouse falls within the nutrition context and consists of clinical records – referring to evaluation records of nutrition episodes – extracted from a Portuguese hospital, recording a period between 1st August, 2011 to 4th January, 2017. The raw data consisted of 2892 medical records and 15 attributes. It is relevant to mention that it was necessary to remove all the rows that had cells with empty or unknown values. After deleting these rows, the final dataset consisted of 1825 records. A meticulous and careful analysis allowed the selection of 6 attributes. These attributes contained information about the patient's characteristics, namely the patient's *Age*, *Weight*, *Height*, *BMI* and its *Nutrition Classification*. In addition, information about the nutritionist's answer concerning if the patient will or not be accompanied by the nutrition service was also added– *NutriFollow-Up*. The dataset values are numeric or nominal and discrete or continuous according to the nature of the attribute. Thus, all attributes were defined as numeric and continuous, except the *Nutrition Classification* and the target attribute value, *NutriFollow-Up,* which were defined as nominal and discrete since the possible values were already pre-defined corresponding respectively to {underweight, normal weight, pre-obesity, obesity class I, obesity class II, obesity class III} and {no, yes}. It's important to note that after the normalization those ranges changed to {0, 0.2, 0.4, 0.6, 0.8, 1} and {0, 1}.

B. Data preparation and Data Modeling

The data preparation stage covers all the steps performed to construct and prepare the raw data into the final dataset in order to be fed into the data modeling phase [15]. Since the data used in this project came from a previously developed data warehouse, data transformation from the data warehouse relational structure, with its multiple tables, to a form suitable for data mining was a crucial step. Data mining algorithms are usually based on a single table, within which there is a record for each individual, and the fields contain variable values specific to the individual. The most portable format is a flat file, with one line for each individual record. This flat file is created by one or more Structured Query Language (SQL in advance*)* statements on the data warehouse. In this sense, data was extracted from the *MySQL Workbench* database into an *Excel* file which, then, was subject to data cleaning and transformation procedures.

As mentioned, the data warehouse consisted of clinical records and consequently was filled by health professionals. The lack of standards and rules for filling those records lead to a high probability of errors and irregularities. The data cleaning and preprocessing goal is to transform the dataset by removing inconsistencies, noise, bias, incoherence, and redundancies characterizing medical data [15]. Data mining is a naturally iterative process, where several steps need to be repeated several times. With this in mind, although some data cleaning procedures have been already performed on the data warehouse, additional cleaning procedures were still required, and data preprocessing was performed several times in order to find the dataset with the highest accuracy. In this sense, a search for errors, missing values and inconsistencies that may compromise data integrity was conducted. After an exhaustive and intensive examination, numerous errors were found, namely blank spaces and information with writing errors and symbols. As mentioned, the rows that had columns with empty or unknown values were removed. The fields with inconsistencies were standardized by removing unexpected symbols and units of measures and by the transformation of all the values in the same unit of measure. Additionally, the *Nutrition Classification* values "no" and "yes" were converted to Boolean values (0 and 1). Similarly, the *NutriFollow-Up* values were converted to Boolean and changed from {underweight, normal weight, pre-obesity, obesity class I, obesity class II, obesity class III} to {0, 0.2, 0.4, 0.6, 0.8, 1}. Subsequently, the data in the *Excel* file was submitted to several data transformations. These transformations include: Normalization, where all the data values were sized to fall within a small range (0–1) by setting the min value to 0 and max value to 1; Smoothing, that performs a search for the occurrence of values out of range (noise values) and removes them from the data; And Discretization, which splits the range of continuous attributes into intervals.

In order to be readable by the data mining software, the dataset was saved in Comma-Separated Value (CSV in advance*)* format, which is a *Weka's* supported format. Although *Weka* accepts files in CSV format, it expects data to be in its own Attribute-Relation File Format (ARFF in advance*)* because it's necessary to have type information about each attribute, which cannot be automatically deduced from the attribute values [16]. This format is an extension of the CSV file format, where a header is used to provide metadata about the data types in the columns. In this sense, by clicking in the *Weka* "*Tools*" menu and selecting the "*ArffViewer*", the CSV file was loaded and then saved in ARFF. Once the data is precisely recorded, the ARFF file is loaded into *Weka*, and the classifiers can be applied.

Before the modelling phase can begin, it is necessary to decide which strategy will be used to split the data set into the learning and validation set. A common approach is to learn from two-thirds of the dataset and then to test on the remaining one-third of the sample. Such strategy may not apply to a small dataset, since the learning algorithms may have issues due to the small amount of data for learning, while the test may be still lacking to achieve the desired confidence. A contemporaneous approach to solve this problem, and the one used in this project, is k-fold cross-validation. In this study, the data was divided into 10 data subsets containing almost an equal number of data instances and approximately matching the outcome distribution of the learning set. Then, data from the nine subsets was used for modeling and the remaining subset was used to test the resulting model and assess statistics. The process of training and testing was repeated ten times, each time using a different testing subset [17]. Finally, it was imperative to apply different modeling techniques to the data set. The tested algorithms were Weka classification algorithms and, in order to determine which was the most appropriate algorithm and the one that performed best, several performance evaluation measures were used. Those measures are described in the next sub-section.

C. *Performance Evaluation*

The evaluation of predictive models is made based upon two measures: their predictive performance and comprehensibility. Comprehensibility is a subjective measure that is hard to quantify since it must be evaluated by domain experts. Predictive performance is easier to quantify, and regular statistics comprise sensitivity, specificity, classification accuracy and area under the Receiver Operating Curve (ROC in advance) [17]. In this sense, the classifiers applied to the nutritional dataset will be evaluated based on a certain set of performance measures that include the confusion matrix, accuracy, error rate, recall/sensitivity, precision/specificity, ROC area and Kappa statistic. Accordingly, the best performing classifier is chosen based on the following measures.

The Accuracy of a classifier is the percentage of test set tuples that are correctly classified by the classifier [18]. In contrast, the error-Rate of a classifier, M, is measured as 1-Acc (M), where Acc (M) is the accuracy of M [18].

The sensitivity and specificity measures can be used to calculate accuracy of classifiers. Sensitivity is also referred to as the true positive rate, i.e., the proportion of positive tuples that are correctly identified, while specificity is the true negative rate, which is the proportion of negative tuples that are correctly identified [19] [20] [21].

Confusion Matrix is a specific table layout useful to visualize the classifier's performance by analyzing how well it recognizes the tuples of different classes. It provides information to determine how well the model performed by computing the Accuracy for correct predictions and Error Rate for incorrect predictions [18] [21]. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it is easy to see if the system is confusing two classes [21]. In addition, the confusion matrix does not only show how well the model predicts, but also the details that might go wrong during the data mining process [18].

The ROC area ranges from 0.5 to 1. A classifier with a ROC area of 0.5 is a poor classifier, roughly equivalent to flipping a coin to decide the class membership. The closer the ROC area to 1, the higher the discriminating power of the classifier [22] [18]. Finally, the Kappa statistic reflects the performance and accuracy of the classifier [22].

III. Experimental Results

When testing the original dataset, the performance results obtained by all the classification algorithms were far below expectations. Despite having a relatively high accuracy value, around 80%, the results presented quite high error values. This is most probably due to the fact that, in this case study, as in many others, the values that the class can adopt don't present equal probabilities. In fact, the dataset has a quite abnormal class distribution, in which of the 1825 records, only 345 records are referring to patients who were assigned to be followed by a nutrition specialist. This means that the training set is almost completely constituted by patients who weren't assigned to follow the nutrition service, so, when performing, the learning model is unable to correctly predict the patients who were assigned to follow the nutrition service. Since this imbalance adversely affects the performance results, it was fundamental to study and adopt approaches to balance the dataset. Thereby, different approaches and scenarios were taken into consideration:

TABLE I. THE DIFFERENT SCENARIOUS USED IN THIS STUDY

Data set	Attributes	Balancing Technique	Age grouped in classes
DS1	All attributes	*undersampling*	-
DS2	All attributes	*undersampling*	✓
DS3	Without *Height*	*undersampling*	-
DS4	All attributes	*oversampling*	-
DS5	All attributes	*oversampling*	✓
DS6	Without *Height*	*oversampling*	-

In data analysis, *oversampling* and *undersampling* are procedures adopted to adjust the dataset class distribution, i.e., the proportion of the distinct target categories. These are contrary but approximately equivalent techniques, since both resort to a bias to select more samples from one class than from another [23] [24]. In this sense, both techniques were applied in order to improve the learning model's performance.

A simple *undersampling* technique is to randomly drop some instances from the dominant class, the one with the upper distribution, in this case the patients who weren't assigned to follow the nutrition service, to obtain a balanced dataset of 690 samples with 50% of each class (DS1). In this sense, after preparing the new dataset, several classification algorithms were applied. It is worth to mention the single rule operation of the most primitive learning scheme in Weka, *ZeroR*. Given a new instance for classification, *ZeroR* predicts the majority class in the training data for problems with a categorical class value, and the average class value for numeric prediction problems. This is useful to generate a baseline performance that other learning schemes are compared to [25].

The data mining models included in this study are Decision Trees and Decision Rules. Additionally, in an attempt to increase performance by reducing variance, *Bagging* was applied to the *REPTree* classifier. The performance of the best three algorithms, namely *OneR*, *REPTree* and *Bagging*, on dataset 1 is below presented. *OneR* uses the minimum-error attribute for prediction, discretizing numeric attributes. In turn, *Bagging* is a classifier used to reduce variance based on the *REPTree* algorithm which is a fast decision tree learner.

TABLE II. PERFORMANCE EVALUATION MEASURES OBTAINED ON DS1

Classifier	Accuracy	Error Rate	Specificity	Sensitivity	Kappa statistic	ROC Area
ZeroR	49.2754%	0.5072	0.493	0.493	-0.0145	0.493
OneR	90.1449%	0.0986	0.908	0.901	0.8029	0.901
REPTree	90.8696%	0.0913	0.920	0.909	0.8174	0.915
Bagging	90.4348%	0.0956	0.916	0.904	0.8087	0.937

OneR

```
=== Confusion Matrix ===

   a   b   <-- classified as
 333  12 |   a = 0
  56 289 |   b = 1
```

REPTree

```
=== Confusion Matrix ===

   a   b   <-- classified as
 342   3 |   a = 0
  60 285 |   b = 1
```

Bagging

```
=== Confusion Matrix ===

   a   b   <-- classified as
 341   4 |   a = 0
  62 283 |   b = 1
```

Fig. 1. Confusion Matrix representation of the best three algorithms on DS1.

Subsequently, the numeric data was minimized, in an attempt to increase the classifier's performance, specifically the *Age* data. This transformation was made by classifying the age values into well-defined categories, namely children, teenagers, adults, and elderly (DS2). The performance results for the dataset 2 are below presented.

TABLE III. PERFORMANCE EVALUATION MEASURES OBTAINED ON DS2

Classifier	Accuracy	Error Rate	Specificity	Sensitivity	Kappa statistic	ROC Area
ZeroR	49.2754%	0.5072	0.493	0.493	-0.0145	0.493
OneR	90.1449%	0.0986	0.908	0.901	0.8029	0.901
REPTree	90.7246%	0.0928	0.917	0.907	0.8145	0.924
Bagging	90.5797%	0.0942	0.916	0.906	0.8116	0.932

Through attribute weighting algorithms, it was possible to evaluate the importance and the contribution of each attribute in building the target variable (the nutritionist's response to the need for patient follow-up by the nutrition service). Data was normalized to give a value between 0 and 1 to each weight. In this sense, the attribute evaluator *CorrelationAttributeEval*, which evaluates the worth of an attribute by measuring the correlation between it and the target class, *NutriFollowUp*, was applied to the nutrition dataset with *undersampling* technique (DS1). The results are presented in Fig. 2.

```
Ranked attributes:
 0.65567  5 Nutrition Classification
 0.54099  4 BMI
 0.48781  2 Weight
 0.00727  1 Age
 0.00722  3 Height
```

Fig. 2. Representation of the ranked attributes when applying the attribute evaluator *CorrelationAttributeEval.*

The attribute with less worth to the *NutriFollowUp* prediction is the patient's height. This attribute was removed (DS3) in order to see if the performance of the classifiers increases. The results for dataset 3 are shown below. It is important to refer that the attribute that clearly has more influence on the results is the one related to the patient's nutrition classification.

TABLE IV. PERFORMANCE EVALUATION MEASURES OBTAINED ON DS3

Classifier	Accuracy	Error Rate	Specificity	Sensitivity	Kappa statistic	ROC Area
ZeroR	49.2754%	0.5072	0.493	0.493	-0.0145	0.493
OneR	90.1449%	0.0986	0.908	0.901	0.8029	0.901
REPTree	90.8696%	0.0913	0.920	0.909	0.8174	0.923
Bagging	90%	0.1	0.910	0.900	0.8	0.941

OneR

```
=== Confusion Matrix ===

   a   b   <-- classified as
 333  12 |   a = 0
  56 289 |   b = 1
```

REPTree

```
=== Confusion Matrix ===

   a   b   <-- classified as
 342   3 |   a = 0
  60 285 |   b = 1
```

Bagging

```
=== Confusion Matrix ===

   a   b   <-- classified as
 337   8 |   a = 0
  61 284 |   b = 1
```

Fig. 3. Confusion Matrix representation of the best three algorithms on DS3.

In addition to *undersampling*, it was applied an *oversampling* technique to balance the dataset's class distribution. In turn, simple *oversampling* selects each instance of the class with the lowest distribution, in this case, it is the class related to the patients who were assigned to follow the nutrition service, and copies it to obtain a balanced dataset of 2960 samples with 50% of each class (DS4).

The performance of the best three algorithms applied to dataset 4, namely *RandomForest*, *RandomizableFilteredClassifier* and *RandomCommittee*, is presented below. *RandomForest* constructs a forest of random trees. The *RandomizableFilteredClassifier* is a simple variant of the *FilteredClassifier* algorithm that implements the Randomizable interface, useful for building ensemble classifiers using the *RandomCommittee* meta learner which, in turn, builds an ensemble of randomizable base classifiers.

TABLE V. PERFORMANCE EVALUATION MEASURES OBTAINED ON DS4

Classifier	Accuracy	Error Rate	Specificity	Sensitivity	Kappa statistic	ROC Area
ZeroR	50%	0.50	0.250	0.500	0	0.500
Random Forest	88.8176%	0.1118	0.905	0.888	0.7764	0.962
Random Committee	90.5405%	0.0946	0.916	0.905	0.8108	0.962
Randomizable FilteredClassifier	88.3108%	0.1169	0.900	0.883	0.7662	0.859

RandomForest

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1154  326 |   a = 0
   20 1460 |   b = 1
```

RandomCommitt

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1220  260 |   a = 0
   20 1460 |   b = 1
```

RandomizableFiltered Classifier

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1164  316 |   a = 0
   15 1465 |   b = 1
```

Fig. 4. Confusion Matrix representation of the best three algorithms on DS4.

Similar to what was previously done, the *Age* was grouped in well-defined categories namely children, teenagers, adults, and elderly (DS5), in order to understand how the reduction of numeric data combined with the oversampling technique influences the performance values. The results of the dataset 5 are presented below.

TABLE VI. PERFORMANCE EVALUATION MEASURES OBTAINED ON DS5

Classifier	Accuracy	Error Rate	Specificity	Sensitivity	Kappa statistic	ROC Area
ZeroR	50%	0.50	0.250	0.500	0	0.500
Random Forest	79.3581%	0.2064	0.808	0.794	0.5872	0.872
Random Committee	79.4257%	0.2057	0.804	0.794	0.5885	0.871
Randomizable FilteredClassifier	77.4662%	0.2253	0.789	0.775	0.5493	0.830

The attribute evaluator *CorrelationAttributeEval* was applied to the nutrition dataset with the *oversampling* technique (DS4), in order to figure out which are the most relevant attributes in this setting. The results are shown below.

```
Ranked attributes:
 0.08619  1 Age
 0.04968  4 BMI
 0.04166  2 Weight
 0.03931  5 Nutrition Classification
 0.00902  3 Height
```

Fig. 5. Representation of the ranked attributes when applying the attribute evaluator *CorrelationAttributeEval.*

The attribute with less worth to the *NutriFollowUp* prediction is the patient's height. Once again, this attribute was removed (DS6) in order to see if the performance of the classifiers increases. The results of dataset 6 are shown below.

TABLE VII. PERFORMANCE EVALUATION MEASURES OBTAINED ON DS6

Classifier	Accuracy	Error Rate	Specificity	Sensitivity	Kappa statistic	ROC Area
ZeroR	50%	0.50	0.250	0.500	0	0.500
Random Forest	89.1554%	0.1084	0.907	0.892	0.7831	0.955
Random Committee	90.4392%	0.0956	0.916	0.904	0.8088	0.957
Randomizable FilteredClassifier	87.4662%	0.1253	0.892	0.875	0.7493	0.855

RandomForest

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1174  306 |    a = 0
   15 1465 |    b = 1
```

RandomCommitte

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1217  263 |    a = 0
   20 1460 |    b = 1
```

Randomizable Filtered Classifier

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1139  341 |    a = 0
   30 1450 |    b = 1
```

Fig. 6. Confusion Matrix representation of the best three algorithms without the Height attribute.

IV. Discussion and Conclusions

In this paper, it is shown the role of data mining classification algorithms for the use of evidence-based medicine in the context of nutrition evaluation. The prediction's target value is the nutritionist's response to the need for patient follow-up. The original dataset consists of 1825 records with 6 attributes which comprised the patient's characteristics, namely *Age*, *Weight*, *Height*, *BMI* and its *Nutrition Classification*, and the nutritionist's answer concerning if the patient needs to be followed by a nutritionist – *NutriFollow-Up*. The results obtained with the original dataset weren't favorable. Thus, several changes were made to this dataset, and different approaches and scenarios were taken into consideration (Table 1).

In the first dataset (DS1), the algorithm with the best accuracy is the *REPTree,* since it presents the highest percentage of correctly classified instances, 90.8696%. This means that 627 instances out of 690 were correctly classified. Consequently, it is also the classifier with the smallest error, about 0.0913, which means that only 63 instances were misclassified. This algorithm also holds the best values of specificity (0.920), sensitivity (0.909) and kappa statistic (0.8174), which reflects the performance and accuracy of the classifier. In addition, the best confusion matrix is also the one obtained with the *REPTree* classifier. From Fig. 1, the numbers 342 and 285 indicate the number of cases where the actual and predicted values are the same. In other words, the diagonal shows all the correct predictions. While the number 3 represents the number of cases where the actual outcome was for the patient to not follow the nutrition service, but it was predicted to be assigned to follow the nutrition service, and the number 60 represents the number of cases where the outcome was for the patient to be assigned to follow the nutrition service, but it was predicted to not follow the nutrition service. This shows that, when applying the *REPTree* algorithm to the dataset with the *undersampling* technique, the number of wrong predictions is actually very small. Although the ROC area of the *REPTree* algorithm, 0.915, is a little bit lower than the *Bagging*, 0.937, the best classifier is the *REPTree* since it has the best performance in all the other evaluation measures and its ROC area still indicates that the classifier has a high validity (0.915 is closer to 1).

When grouping the *Age* into classes (DS2), the *ZeroR* and *OneR* algorithms are not affected and the results worsen for the *REPTree* algorithm. Although there is an improvement for the *Bagging* algorithm, the results obtained previously by the *REPTree* algorithm on DS1 remain preferable.

On the other hand, when removing the *Height* attribute (DS3), the *ZeroR* and *OneR* algorithms are not affected and the performance results worsen for the *Bagging* algorithm. Although, the confusion matrix distribution and the values of performance measures remain the same for the *REPTree* algorithm, the ROC area value increases, which means that the obtained results overcome the previously best results. Thus, the *Height* attribute does not substantially influence results, which is unexpected, since it would be expected that the patient's height had influence on determining if he should or not be followed by a nutrition specialist. In conclusion, this means that when applying the *undersampling* technique, the classifier with best results is the *REPTree* and its performance can be improved by removing the Height attribute, reaching an accuracy value of 90.8696%.

In what concerns the first dataset of the *oversampling* technique (DS4), the algorithm with the best accuracy is the *RandomCommittee,* since it presents the highest percentage of correctly classified instances, 90.5405%, which means that 2680 instances out of 2960 were correctly classified. Consequently, it is also the classifier with the smallest error, about 0.0946 and therefore only 280 instances were misclassified. This algorithm also holds the best values of specificity (0.916), sensitivity (0.905), kappa statistic (0.8108) and ROC area (0.962). In addition, the best confusion matrix is also the one obtained with the *RandomCommittee* classifier.

When grouping the *Age* into classes (DS5), the *ZeroR* is not affected and all the performance measures worsen for all the remaining classifiers.

Finally, when removing the *Height* attribute (DS6), the *ZeroR* is not affected and the performance results worsen for both *RandomizableFilteredClassifier* and *RandomCommittee* algorithms. In contrast, all the performance measures improve for the *RandomForest,* except for the ROC area value. However, when comparing these results with the original *oversampling* dataset (DS4), neither the results obtained for the performance evaluation measures nor the confusion matrix prove to be preferable. This means that when applying the *oversampling* technique, the classifier with best performance is the *RandomCommittee* with an accuracy value of 90.5405%. As discussed, in order to obtain promising results, several changes were made to the original data set, which could probably have been avoided if the amount of data constituting the original data set was greater and the class distribution was more or less equal.

This study's main purpose was to create useful models capable of correctly predict the need for a patient to be followed by a nutrition specialist. As already mentioned, two different approaches were used to balance the dataset class distribution and consequently improve the achieved results. From the previous results analysis, both *undersampling* and *oversampling* have largely increased the performance of the classifiers. However, the results presented when applying the *undersampling* technique turned out to be slightly better. The best constructed model was obtained by the *REPTree* algorithm and certified by different metrics reaching a level of accuracy of 91%, a level of specificity of 92%, sensitivity of 91%, precision Kappa statistic of 82%, ROC area of 0.9 and error rate of 9.1%. A higher error is associated with a classification more distanced from the reality. Thus, it's extremely important to achieve lower values of error since we are dealing with a context that is directly associated with the patient's lives. In this sense, the solution provided can be used to support healthcare providers in the decision-making, improving the nutritional condition of the population by allowing to predict patients' outcomes in the context of nutrition evaluation.

Acknowledgement: This work has been supported by Compete: POCI-01-0145-FEDER-007043 and FCT within the Project Scope UID/CEC/00319/2013.

REFERENCES

[1] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, 2017.

[2] M. Esteves, F. Miranda, J. Machado, and A. Abelha, "Mobile Collaborative Augmented Reality and Business Intelligence: A System to Support Elderly People Self-care", in Advances in Intelligent Systems and Computing, Springer, 2018 (accepted).

[3] A. K. Sigurdardottir, H. Jonsdottir, and R. Benediktsson, "Outcomes of educational interventions in type 2 diabetes: WEKA data-mining analysis," Patient education and counseling, 2007, pp.21-31.

[4] S. Oliveira, F. Portela, MF. Santos, J. Machado, A. Abelha, A. Silva, and F. Rua, "Clustering data mining models to identify patterns in weaning patient failures," International journal of biology and biomedical engineering, 2016.

[5] H. Singh, and K. S. Kaswan, "Clinical decision support systems for heart disease using data mining approach,".

[6] D. Li, H. W. Park, E. Batbaatar, Y. Piao, and K. H. Ryu. "Design of health care system for disease detection and prediction on hadoop using DM techniques," Conf. Health Informatics and Medical Systems, 2016.

[7] Ribeiro, F. Portela, M. F. Santos, J. Machado, A. Abelha and F. R. Martins. "Predicting Patients admission in Intensive Care Units using Data Mining". In Research Jornal POLIBITS, 2016. (accepted for publication).

[8] R. Peixoto, F. Portela, M. F. Santos, A. Abelha, J. Machado and F. R. Martins. Predicting resurgeries in Intensive Care using Data Mining. 16th International Conference on Biomedical Engineering (ICBME 2016). Singapore. Springer, IFMBE, 2016.

[9] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," International Journal on Computer Science and Engineering, 2010, pp.250-255.

[10] R. Vidya, and G. M. Nasira, "Knowledge extraction in medical data mining: a case based reasoning for gynecological cancer an expert diagnostic method," 2006.

[11] H. Peixoto, J. Machado, J. Neves and A. Abelha. "Semantic Interoperability and Health Records". In E-Health. Springer, Berlin, Heidelberg, 2010. p. 236-237.

[12] S. G. Jacob, and R. G. Ramani, "Mining of classification patterns in clinical data through data mining algorithms," In Proceedings of the International Conference on Advances in Computing, Communications and Informatics, ACM, August 2012, pp.997-1003.

[13] A. Ribeiro, F. Portela, M. Santos, A. Abelha, J. Machado, and F. Rua, "Patients' admissions in intensive care units: a clustering overview," Information, 2017, pp.23.

[14] Reis, R., Mendonça, A., Ferreira, D. L. A., Peixoto, H., & Machado, J. (2017) "Business Intelligence for Nutrition Therapy". In Next-Generation Mobile and Pervasive Healthcare Solutions (pp. 203-218). IGI Global.

[15] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, and A. Geissbuhler, "Clinical data mining: a review," Yearb Med Inform, 2009, pp.121-133.

[16] I. Witten, and E. Frank, "Weka machine learning algorithms in java," 2000.

[17] R. Bellazzi, and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," International journal of medical informatics, 2008, pp.81-97.

[18] D. H. Qudsi, M. Kartiwi, and N. B. Saleh, "Predictive data mining of chronic diseases using decision tree: a case study of health insurance company in Indonesia," International Journal of Applied Engineering Research, 2017, pp.1334-1339.

[19] A. Mani, et al., "Data mining strategies to improve multiplex microbead immunoassay tolerance in a mouse model of infectious diseases," PloS one, 2015.

[20] A. V. Kumar, R. F. Ali, Y. Cao, and V. V. Krishnan, "Application of data mining tools for classification of protein structural class from residue based averaged NMR chemical shifts," Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2015, pp.1545-1552.

[21] A. H. M. Ragab, A. Y. Noaman, A. S. Al-Ghamdi, and A. I. Madbouly, "A comparative analysis of classification algorithms for students college enrollment approval using data mining," Workshop on Interaction Design in Educational Environments, ACM, June 2014, p.106.

[22] J. M. Hardin, and D. C. Chhieng, "Data mining and clinical decision support systems," Clinical Decision Support Systems, Springer, 2007.

[23] N. V. Chawla, "Data mining for imbalanced datasets: An overview," Data mining and knowledge discovery handbook, Springer US, 2005, pp. 853-867.

[24] M. M. Rahman, and D. N. Davis, "Addressing the class imbalance problem in medical datasets," International Journal of Machine Learning and Computing, 2013, p. 224.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter, 2009, pp. 10-18.