Soft Computing (2022) 26:12041–12055
https://doi.org/10.1007/s00500-022-07503-z

Machine learning techniques for prediction of multiple sclerosis progression

Dario Branco[1] · Beniamino di Martino[1,2] · Antonio Esposito[1] · Gioacchino Tedeschi[3] · Simona Bonavita[3] · Luigi Lavorgna[3]

Accepted: 2 September 2022 / Published online: 22 September 2022

Abstract

Patients afflicted by multiple sclerosis experience a relapsing-remitting course in about 85% of the cases. Furthermore, after a 10/15-year period their situation tends to worse, resulting in what is considered the second phase of multiple sclerosis. While treatments are now available to reduce the symptoms and slow down the progression of the disease, the administration of drugs must be adapted to the course of the disease, and predicting relapsing periods and the worsening of the symptoms can greatly improve the outcome of the treatment. For this reason, indicators such as the patient-reported outcome measures (PROMs) have been largely used to support early diagnosis and prediction of future relapsing periods in patients affected by multiple sclerosis. However, such indicators are insufficient, as the prediction they provide is often not accurate enough. In this paper, machine learning techniques have been applied to data obtained from clinical trial, in order to improve the prediction capabilities and provide doctors with an additional instrument to evaluate the clinical situation of patients. After the application of correlation indicators and the use of principal component analysis for the reduction of the dimensionality of the feature space, classification algorithms have been applied and compared, in order to identify the best suiting one for our purposes. After the application of re-balance algorithms, the accuracy of the machine learning-based prediction system reaches 79%, demonstrating the capability of the framework to correctly predict future progression of disability.

Keywords Multiple sclerosis · Prediction · Classification algorithms · Re-balance techniques · Principal component analysis

1 Introduction

Multiple sclerosis (MS) is a chronic, inflammatory, demyelinating disease of the central nervous system with a variable course. At onset about 85% of MS patients experience a **Relapsing-remitting** (RR) course, with relapses (new focal neurologic signs and symptoms caused by inflammation and demyelination) followed by periods of remission (Confavreux and Vukusic 2006). About 10–15 years later, almost 50% of RRMS patients develop a progressive phase with insidious worsening independent from relapses that is the secondary progressive phase of MS (SPMS) (Grothe et al. 2016). RRMS and SPMS responses to available treatments are significantly different therefore adapting treatment to the phase of disease is critical for patient outcomes. Recently, siponimod was approved as it was found to slow disability accumulation compared with placebo in patients with SPMS,

Simona Bonavita and Luigi Lavorgna have contributed equally to this work.

✉ Beniamino di Martino
beniamino.dimartino@unicampania.it

Dario Branco
dario.branco@unicampania.it

Antonio Esposito
antonio.esposito@unicampania.it

Gioacchino Tedeschi
gioacchino.tedeschi@unicampania.it

Simona Bonavita
simona.bonavita@unicampania.it

Luigi Lavorgna
luigi.lavorgna@unicampania.it

[1] Department of Engineering, University of Campania “Luigi Vanvitelli”, Aversa, CE, Italy

[2] Department of Computer Science and Information Engineering, Asia University, Taichung City, Taiwan

[3] Department of Advanced Medical and Surgical Sciences, University of Campania “Luigi Vanvitelli”, Caserta, Italy

especially in those with active disease. In this perspective early identification of the transitioning phase from the RR to the SP phase is mandatory to adapt disease modifying treatment.

Factors predicting progression to SPMS in patients with RRMS are not well established; indeed it has been reported that the transition from RRMS to SPMS is characterized by a diagnostic uncertainty that lasts almost 3 years (Rojas et al. 2021). A possible reason for this delay is that indicators of SPMS may be subtle. For example, patients may report a worsening in their condition but the neurologic examination may detect little or no change in their status; as a consequence, patients may keep on therapies ineffective for SPMS, with unnecessary adverse effects and costs.

In 2009, the FDA published guidance on **Patient Reported Outcome Measures** (PROMs), which was defined as "any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else" [ref].

PROM is an umbrella term that includes evaluations of symptoms, health-related quality of life (QoL), health status, depression, well-being, treatment satisfaction, and adherence to treatment. Therefore, PROMs complement and support outcome measures based on clinical assessments. Recently, Giovannoni et al. proposed to incorporate PROMs in the definition of NEDA (**No Evidence of Disease Activity**) status, typically a combination of outcome measures. To avoid missing the therapeutic window for SPMS, it is important to expedite the accurate diagnosis of SPMS enabling appropriate treatment to start in a timely manner.

Traditional clinical measurements, such as the aforementioned PROMs, are not enough accurate to predict secondary progression in MS if taken alone. However, applying machine learning techniques and algorithms to integrate the results obtained through PROMs would surely improve the prediction capability and lead to more reliable results.

In the specific case study addressed by this work, the data obtained from a clinical trial that has involved 220 different patients has been analyzed by means of statistical and machine learning techniques, in order to assess the best course of action to correctly predict the progression of the MS disease.

Indeed, the reduced number of available samples in the original dataset, together with its complexity deriving from the considerable number of features describing each sample, required the application of different data preparation techniques, in order to avoid or at least reduce overfitting and imbalance problems. Similar issues have been addressed in other works, related to different dominions, such as in Di Martino et al. (2021). In particular, since predicting the progression of the MS disease through a single indicator, namely the **Expanded disability Status Scale** (EDSS) described in Sect. 3, proved to be highly impractical due to the nature of the available dataset, the related regression problem was transformed into a simpler binary classification. Binary classification is surely much easier to handle than regression, especially with a very limited dataset as the one currently taken in consideration. However, even with a simplified classification, overfitting errors are bound to arise, when combining the complexity of a dataset with more than 20 features per sample, and a low number of usable samples. The dataset has been thus subjected to a dimensionality reduction, carried out by means of correlation based techniques such as the **Pearson correlation coefficient** (PCC), which was applied with the support of domain experts that confirmed the evaluations that were first obtained through automatic techniques were indeed correct. The dataset with a reduced number of features has been then analyzed through different classifiers, such as the logistic regression, the Gaussian Naive Bayes, the random forest and the linear support vector classification, in order to identify the most accurate one. First, the results showed that, due to the strong imbalance between the two identified classes used in our training, the classifiers generated highly accurate models with very low precision and recall on the less common class.

In order to correct this behavior, re-balance techniques were taken in consideration and applied, and the same classifiers were tested against the new re-balanced dataset. Applying the random oversampling, synthetic minority oversampling technique (SMOTE) and the adaptive synthetic sampling (ADASYN) provided different results, which affected the classifiers differently. By comparing the results, it was evident that using the SMOTE approach with a random forest classifier leads, in our case, to the best accuracy with the lowest overfitting. The final accuracy obtained by the classifier was of 75%, with balanced precision and recall on the examined classes. In agreement with the domain experts these results are satisfactory, the reasons will be investigated in the conclusions. The remainder of this paper is organized as follows: Sect. 3 describes the applied methodology, explaining why the prediction/regression problem was transformed into a binary classification one, and also introduces the used classifiers; Sect. 4 explains how the prediction of the EDSS score was transformed into a binary classification; Sect. 5 explores the dimensionality reduction problem and shows the results obtained with the application of the Pearson correlation technique; Sect. 6 tests the classifiers and compares their results, focusing on the imbalance errors and on their solution; Sect. 6.2 describes the tuning of the random forest classifier though the optimization of its hyperparameters; Sect. 7 closes the paper with final remarks and consideration on the carried out work.

2 Related works

The study of multiple sclerosis outcomes, in particular in regard to relapse-remitting cases, has been carried out in existing works in the literature. A previous study performed on 42 relapsing-remitting MS patients and 30 healthy subjects showed an association of early MS inflammatory disease activity with future clinical disability with an accuracy between 70.6% (for Minimal Activity of Disease Activity, MEDA) and 71.4% (for clinical worsening) (Damasceno et al. 2020). The methodology applied relied on logistic regression, but did not take in consideration other possible machine learning approaches, which could provide better results, especially after the optimization of the features and of the hyperparameters.

In Muthuraman et al. (2020) the authors applied longitudinal 3T MRI and advanced computational models in 2 independent cohorts of patients with early MS (119 patients with early relapsing-remitting MS and a replication cohort of 81 patients) to investigate how white matter (WM) lesion distribution and cortical atrophy topographically interrelate and affect functional disability, predicting individual disability progression with an accuracy of 88% (study cohort) and 89% (replication cohort), respectively. The study focused on cerebral characteristics of patients, rather than on simple questionnaires like in our case, which do not require any complex examination and can be easily administered during routine check-ups.

In another study (Brichetto et al. 2019) machine learning, applied to patient-reported (PROs) and clinical-assessed outcomes (CAOs) to predict disease progression, in a dataset of 3398 evaluations from 810 persons with MS, was able to foresee the course with an accuracy of 82.6%. The main difference between this approach and ours is represented by the dimension of the datasets used to train the algorithms.

3 Overview of the methodology

This study is a secondary data analysis from a previous study reporting PROMs completed by MS patients at their scheduled clinic attendance, between May 2017 and December 2017.

The study was approved by the ethics committee of the University of Campania "Luigi Vanvitelli." Signed informed consent was obtained from each patient prior to enrollment in the study according to the Declaration of Helsinki.

The following questionnaires were administered: The health-related QoL with the 36-item Short-Form Health Survey (SF-36) to measure QoL as an indirect measure of treatment effect. Mental (MCS) and physical (PCS) composite scores ranges from 0 to 100 (Pedregosa et al. 2011), lower scores indicating worse QoL. The Patient-Reported Indices in MS (PRIMUS) (Chawla et al. 2002), a 15-item assessment, to evaluate changes in activities of daily living (PRIMUS activities, score 0–30) and QoL (PRIMUS QoL, score 0–22), higher scores indicating worse activity limitation. The Treatment Satisfaction Questionnaire for Medication (TSQM) (He et al. 2008) to assess treatment satisfaction. It is a 14-item assessment divided in four domains: effectiveness (3 items), side effects (5 items), convenience (3 items) and global satisfaction (3 items).The TSQM-9 domain scores range from 0 to 100 with higher scores representing higher satisfaction on each domain. The Fatigue Severity Scale (FSS), to evaluate fatigue and its effects on daily living. It is a 9-item questionnaire, grading of each item ranges from 1 to 7 (scores > 36 indicating fatigue). The Beck Depression Inventory-II (BDI-II) is used to evaluate depression. It is a 21-item assessment, with different standardized cutoffs for total score to determine the depression's severity: 0–13 no depression, 14–19 mild depression, 20–28 moderate depression, 29–63 severe depression. Disability was assessed by the Expanded disability Status Scale (EDSS) score in 2017 and then retrieved from medical records in 2020 to calculate the ΔEDSS (EDSS 2020-EDSS 2017); progression of the disease was defined when the ΔEDSS > 1 (or half a point if the baseline EDSS score was equal to 5.5).

In order to analyze the data collected through the questionnaires, machine learning techniques have been applied. The objective is to verify the progression of the multiple sclerosis disease by taking in consideration the value of the EDSS at two different times, together with a series of other parameters that are collected through the questionnaires. The structure of the data has a strong influence in the applicable techniques and needs to be considered before the application of any machine learning algorithm.

The available dataset contains 220 different samples, each of which is characterized by 17 different attributes. All attributes are reported in Table 1 and, while being numerical, they have very different ranges.

The EDSS measured at the time of the questionnaire (2017) is part of the available attributes. The measurement of the EDSS after a 3-year period (2020) is also available for each of the samples, but it is not considered as an attribute, as the objective of the study is to predict its value.

It is evident that some critical aspects need to be addressed before any machine learning technique can be applied.

First of all, the number of available samples is quite scarce: predicting the value of the EDSS after a 3-year period by using a regression technique would be rather difficult and would lead to the misleading results. For this reason, instead of trying to predict the exact EDSS value through regression, this study takes in consideration a binary classification problem, in which two classes **Stable** and **Progressed** are taken in consideration. The criteria applied to determine the class to which each sample belongs to are explained in Sect. 4.

Table 1 Numeric data from questionnaires

Name	Description
EDSS (2017)	Expanded Disability Status Scale
PRIMUS tot	Patient Reported Outcome indices for Multiple Sclerosis
BDI tot	A1
TSQM tot	The Treatment Satisfaction Questionnaire For Medication
TSQM effic 1–3	
TSQM coll eff 4–8	
TSQM conv 9–11	
TSQM global sat 12–14	
FSS tot	Fatigue Severity Scale
SF-36 AF	Short Form questionnaire
SF-36 RF	Short Form questionnaire
SF-36 DF	Short Form questionnaire
SF-36 SG	Short Form questionnaire
SF-36 VT	Short Form questionnaire
SF-36 AS	Short Form questionnaire
SF-36 RE	Short Form questionnaire
SF-36 SM	Short Form questionnaire

Another aspect to be considered regards the number of features that describe each sample: Dimensionality reduction is mandatory with such a small dataset, but since there are known connections among the considered features, correlation-based methods are used as described in Sect. 5.

After data has been correctly prepared for ingestion by machine learning algorithms, four classifiers will be taken in consideration in order to verify which one will lead to the best results in terms of accuracy, precision and recall. In particular, these classifiers will be compared:

- A **Logistic Regression**-based classifier. Logistic regression can be used to predict a binary outcome and is known to work quite well with small datasets, so it is a good candidate to test.
- The **Linear Support Vector Classifier** (Linear SVC) is a well-known classifier where a linear model is used to determine a rule that divides members of different classes. Again, the binary nature of the problem we are addressing makes the linear SVC a suitable candidate.
- The **Gaussian Naive Bayes** classifier uses Bayes theory to determine the probability of a tested sample to belong to a specific class, considering absolute independence among the examined samples. Despite the strong assumptions on the input, it often outperforms much more complex classifiers.
- The **Random Forest Classifier** is an ensemble techniques that uses several other predictors, namely a fixed set of decision trees, to calculate a solid classification rule. It is known to be less keen to overfitting, something that should be taken in consideration with a small dataset like ours.

Evaluation of the classifiers will not only have to take in consideration accuracy and overfitting, but also issues with data imbalance, which are very likely to arise (see Sect. 4 for details). To solve this problem, oversampling techniques will be applied and the effect will be described in Sect. 6.1.

4 Definition of the classification problem

This section shows the algorithm in detail for generating the classes that will be used in the classifier. This algorithm is present in the literature and is reported here:

where I is the number of patients and the ith element represents the ith patient. After the execution of the algorithm, a state will be associated for each vector of independent variables (one for each patient) which can be 0 if the disease after 3 years is stable and 1 if over 3 years the disease has worsened. It should be noted that the case of improvement of the patient has not been considered since the cases of improvement in the case of multiple sclerosis are almost nonexistent.

After the application of this algorithm, the dataset contains 54 samples belonging to the Progressed Class, while 166 samples belong to the Stable Class. It is evident that the dataset is imbalanced, as the ration between the Stable Class and Progressed Class is 3:1. The effects of this imbalance will be evident in the preliminary results described in Sect. 6.

Algorithm 1 Class Generation

```
Require: $I > 0$ $EDSS\_old_i \geq 0$ $EDSS\_actual_i \geq 0$
$status_i \leftarrow 0$
$i \leftarrow I$
$delta_i \leftarrow EDSS\_actual_i - EDSS\_old_i$
$y_i \leftarrow EDSS\_old_i$
while $i \neq 0$ do
    if $y_i < 5.5$ and $delta_i \geq 1$ then
        $status_i \leftarrow 1$    ▷ Disease worsening class - Progressed
    else if $y_i < 5.5$ and $delta_i < 1$ then
        $status_i \leftarrow 0$    ▷ Disease stability class - Stable
    else if $y_i \geq 5.5$ and $delta_i \geq 0.5$ then
        $status_i \leftarrow 1$    ▷ Disease worsening class - Progressed
    else if $y_i \geq 5.5$ and $delta_i < 0.5$ then
        $status_i \leftarrow 0$    ▷ Disease stability class - Stable
    end if
    $i \leftarrow i - 1$
end while
```

5 First dimensionality reduction

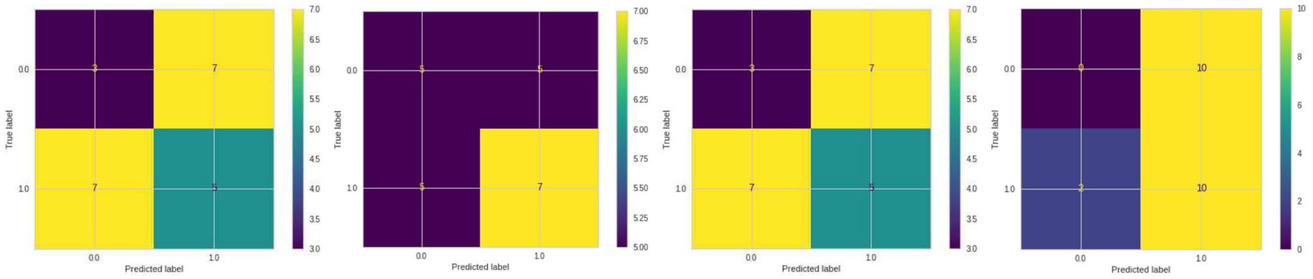
Since the data is not very numerous, a first dimensionality reduction was necessary using the Pearson correlation indices. The Pearson correlation method is a standard approach used to identify correlated features that will not provide substantial information to the dataset and that is especially useful for problems where both input and output data are numeric, as in our original case. Figure 1 shows the obtained results, through confusion matrices, when no feature reduction is applied at all. It is possible to note that logit and random forest algorithms are able to correctly recognize part of the data, with an accuracy of 60%. The Gaussian- and SVC-based approaches do not provide the correct results at all. It is evident that the low number of samples negatively affects the classification, and this fact can only be amplified by the excess of features, which justifies the attempt to obtain better results just by removing statistically irrelevant characteristics.

Since we already had some insight on the feature correlation, thanks to the domain experts who provided the dataset, we used Pearson to confirm their hypothesis and to statistically justify the removal of certain features. The Pearson correlation method is the most common method to use for numerical variables, as testified in Mukaka (2012) Yaping and Changyin (2021) Nasir et al. (2020); it assigns a value between -1 and 1, where 0 is no correlation, 1 is total positive correlation, and -1 is total negative correlation. This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases. Pearson's correlation matrix is provided in Fig. 2.

We decided to eliminate all parameters that have a correlation greater than 0.69 (and not 0.7 as suggested by the Pearson technique) for reasons deriving from the field of application since some parameters are related to each other due to the way in which they were collected. After the dimensionality reduction parameters were reduced to 7, of which the correlation matrix is provided in Fig. 3.

It should be noted that BDI TOT and PRIMUS TOT are present in the data pool despite having a correlation of 0.7 between them for reasons of domain expertise. In support of this dimensional reduction decision, the graph with the feature importance in Fig. 4 with respect to the current EDSS of the chosen independent variables is provided, note how they have almost the same impact on the target value. The feature importance is calculated using a forest of trees, the blue bars are the feature importance of the forest, along with their inter-trees variability represented by the error bars. Feature importance are computed as the mean and standard deviation of accumulation of the impurity decrease within each tree.

After reducing the number of parameters to 7 with the Pearson technique, the principal component analysis (PCA) technique was used to obtain a further dimensionality reduction of the initial dataset, trying not to lose the information contained in the initial dataset. Considering the scarcity in the number of data (which will be treated and solved in the following sections by applying oversampling techniques) it is important to keep the total explained variance as high as possible in order to avoid overfitting problems. For this purpose, the graph of the total explained variance is provided as



(a) Logit (b) Gaussian NB (c) Random Forest (d) Linear SVC

Fig. 1 Confusion matrices for the tested classifiers without any dimensionality reduction

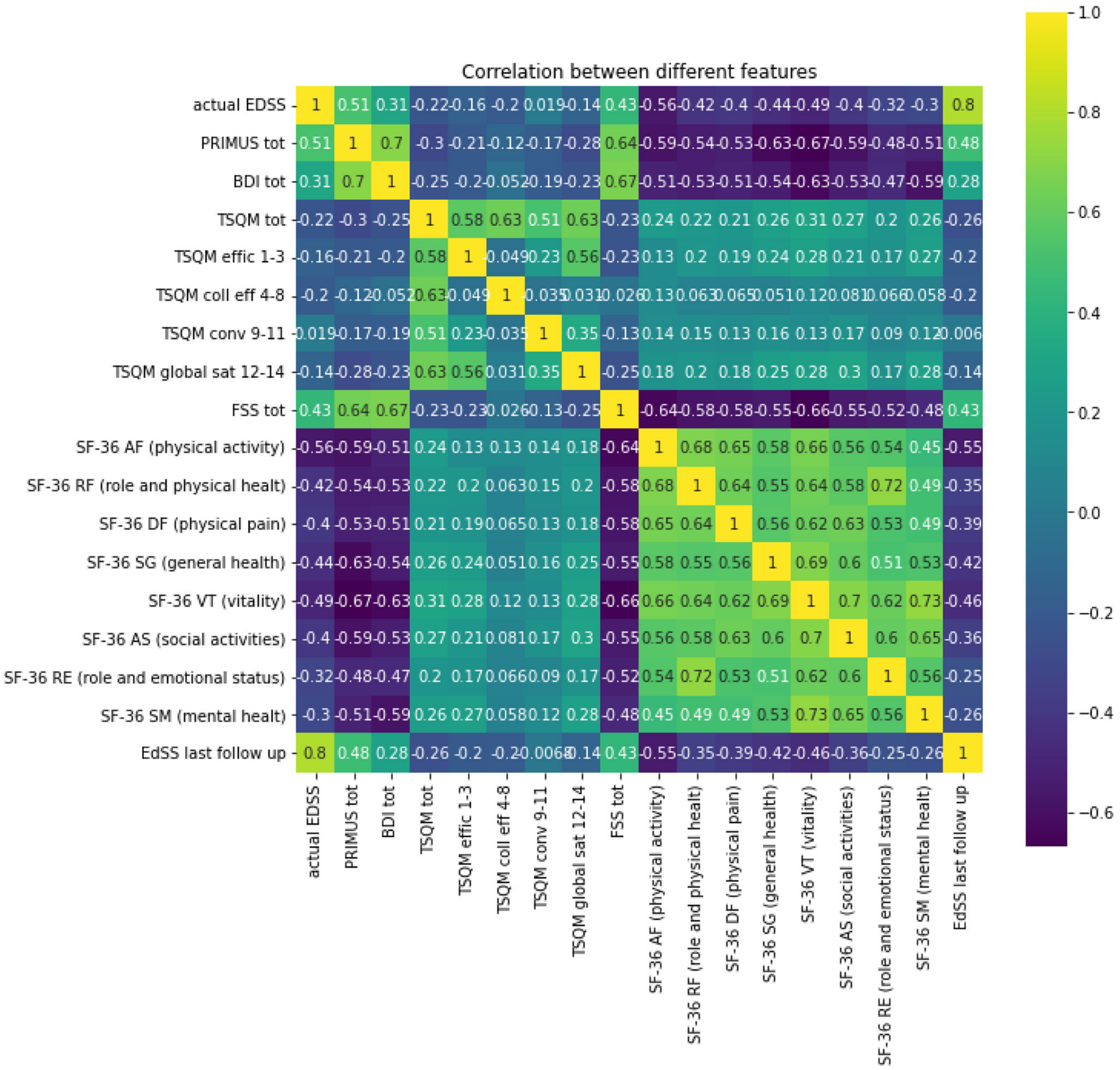


Fig. 2 Pearson correlation view

the number of PCA components varies in Figs. 5 and 6. From the graph, it can be seen that the optimal number of components for dimensional reduction is 5 components since the loss of information is negligible and at the same time we have transformed data from a higher-dimensional space into a lower-dimensional space so that the lower-dimensional representation retains some meaningful properties of the original data (total explained variance equal to **98.3%**). In conclusion, the dimensional space was reduced from 16 dimensions to only 5 dimensions, 9 of which were eliminated by the application of the Pearson correlation algorithm and then the remaining 7 were subjected to linear transformation through PCA for the generation of a new space with only 5 dimensions.

6 Solving scarcity and imbalance of the dataset: classifiers comparison

All the tests on the classifiers were performed by using the Scikit-Learn (Pedregosa et al. 2011) Python library, which offers a wide variety of classification algorithms and other data processing functions. All the classifiers were tested with

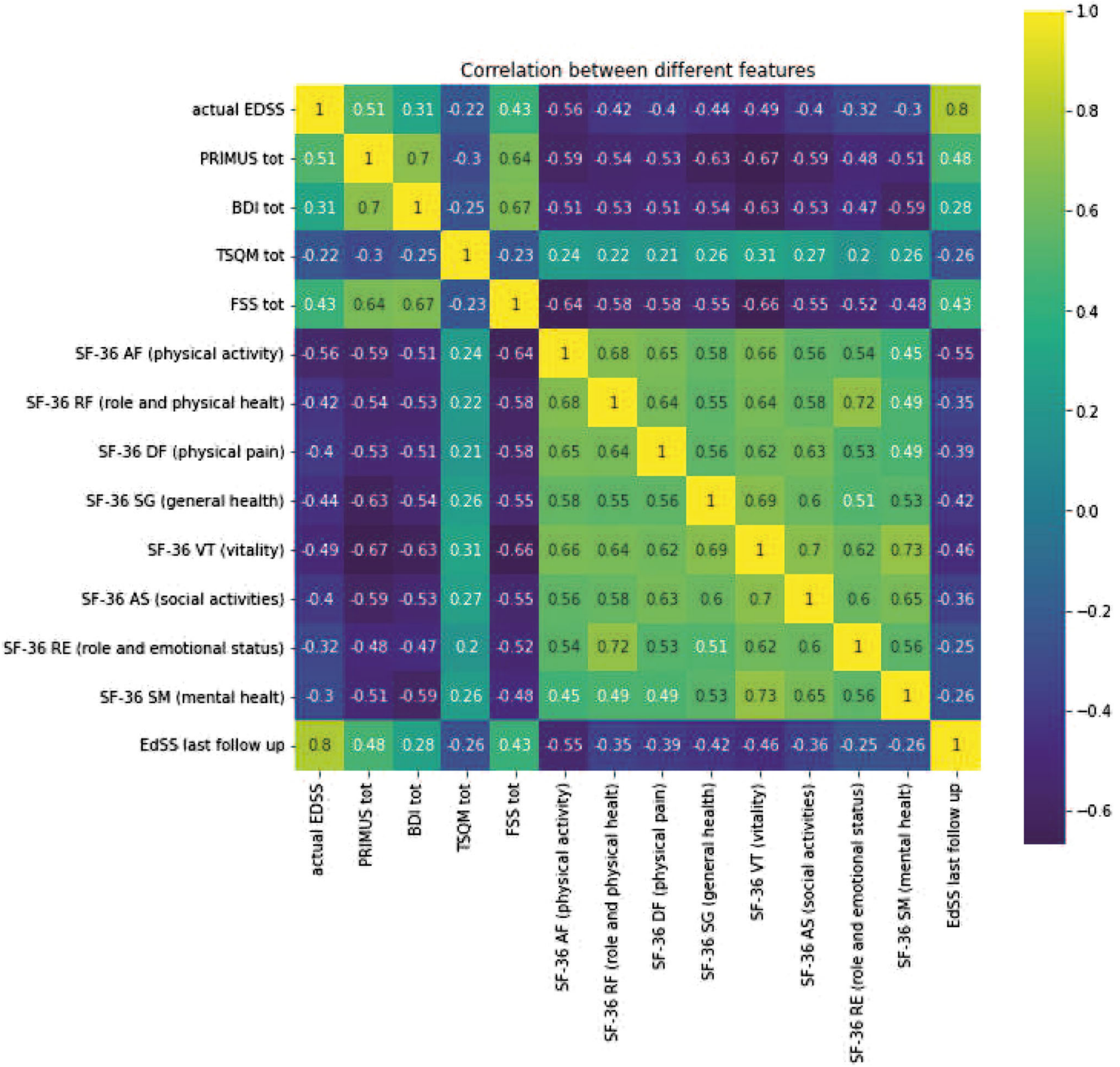


Fig. 3 Pearson correlation view

the default settings, as we only wanted to identify the best one for our case study and then optimize it as shown in Sect. 6.2.

Since the features values strongly varied in range, the dataset has been first normalized by using the Scikit-Learn Standard Scaler. Using the train_test_split function offered by the Scikit-Learn library the dataset was divided into a training set (80% of original data) and test set (20% of original data).

Since the results obtained with the application of PCA, shown in Sect. 5, seemed promising, we tested the different classifiers both with and without its application.

For each of the classifiers, the accuracy was calculated and compared for both the training and test sets, to verify if overfitting occurred. As shown in Fig. 7, test accuracy is above 75% for all classifiers, and the training accuracy does not differ much in the logistic regression, linear SVC and Gaussian NB classifiers, so it seems they do not overfit. Random forest instead performs perfectly on the training set, with a 100% accuracy, while the test accuracy is above 75% as with the other classifiers. In general, all classifiers have been positively influenced from the application of PCA.

While random forest surely overfits our dataset, a closer look to the other scores that have been calculated by Scikit-

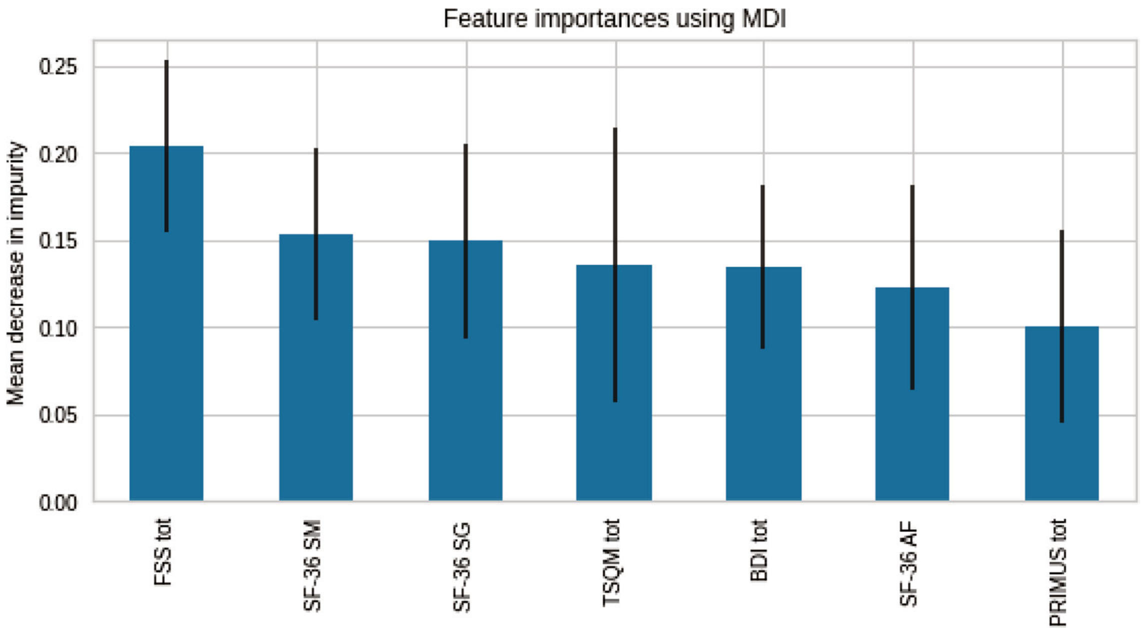


Fig. 4 Random forest feature importance

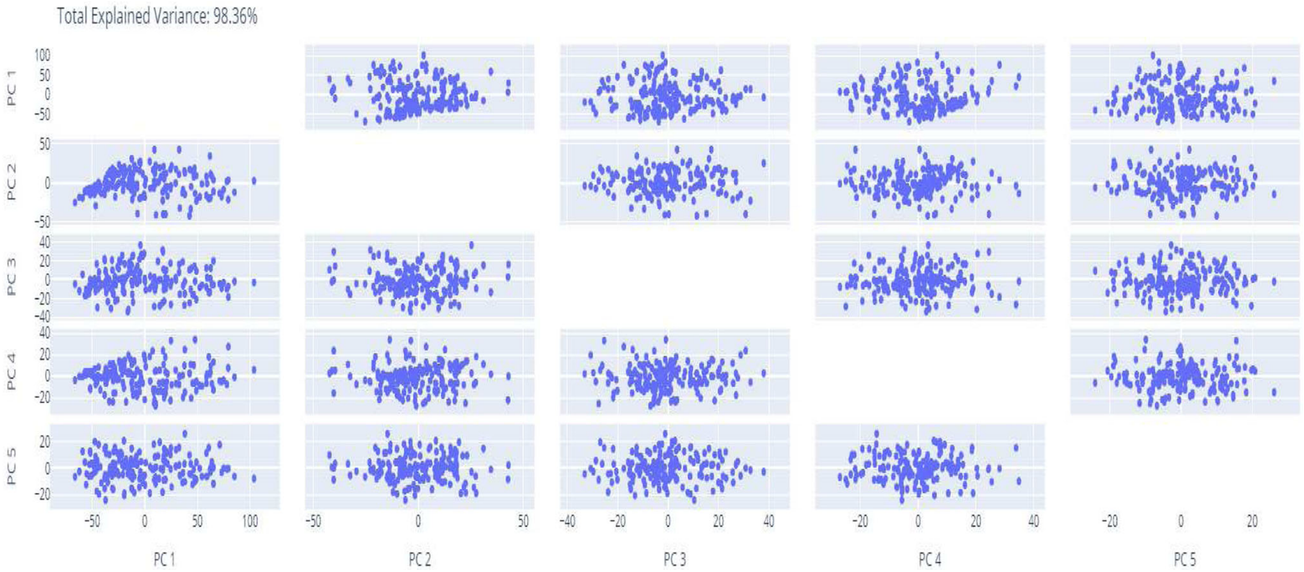


Fig. 5 Five-Component PCA total explained variance

Learn during the training makes it clear that also the other classifiers do not perform well, despite the high accuracy. As shown in Figs. 8 and 9, the confusion matrices calculated for the different classifiers demonstrate that they are not able to recognize the 1.0 class, which corresponds to **Progressed**: They simply always predict the 0.0 class, corresponding to **Stable**, since it covers 75% of the entire dataset.

This situation is even clearer if we look at the Precision, Recall and F1 Score metrics, shown in Table 2. Measures show that, while the classifiers reach a high score in Precision and Recall for the **Stable** class, they have a 0 score when it comes to the **Progressed** class. Only random forest shows a slightly better result, but it is very marginal and, as we have seen before, is obtained through overfitting.

Again, the application of PCA has a very slight positive effect on all classifiers, whose measures do not vary much, apart from the random forest that doubles the precision on the less common class.

6.1 Solving imbalances

As it resulted from the first tests carried out on the dataset, the imbalance between the Stable and Progressed class made the classifiers almost never predicting the Progressed outcome,

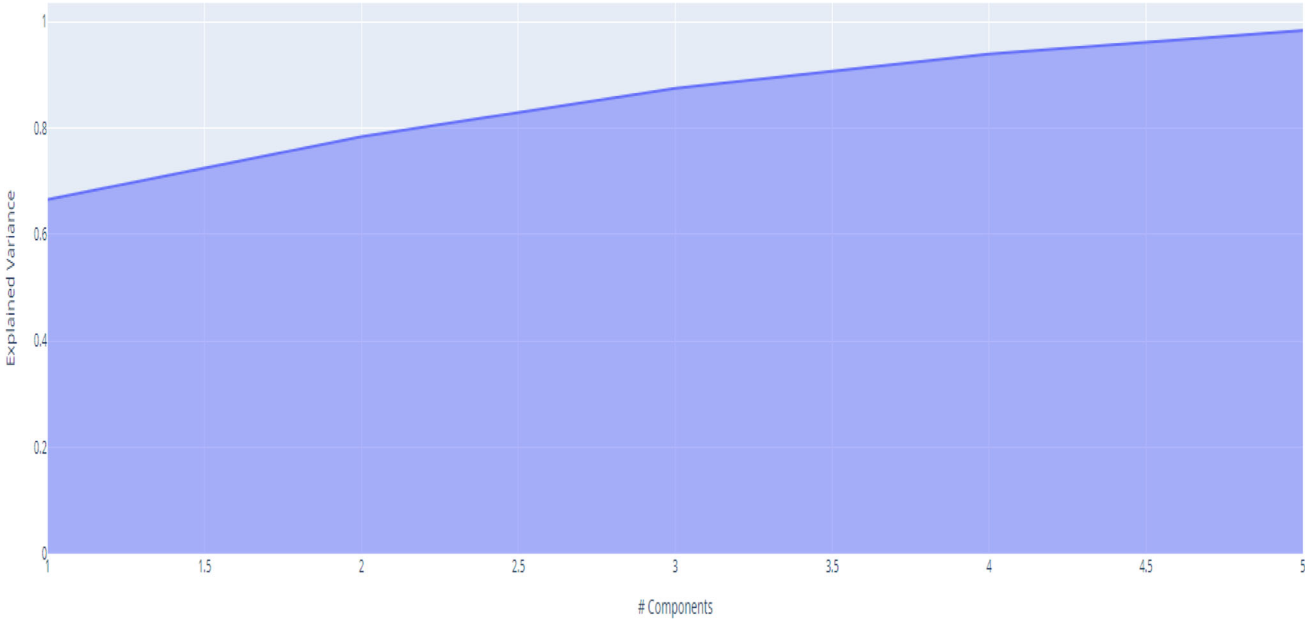


Fig. 6 PCA total explained variance over components

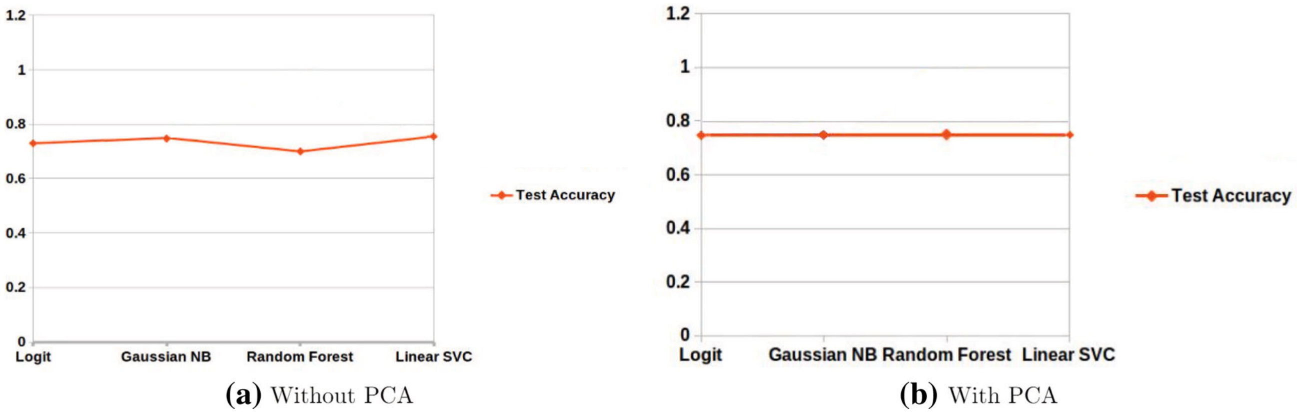


Fig. 7 Comparing training and test accuracies

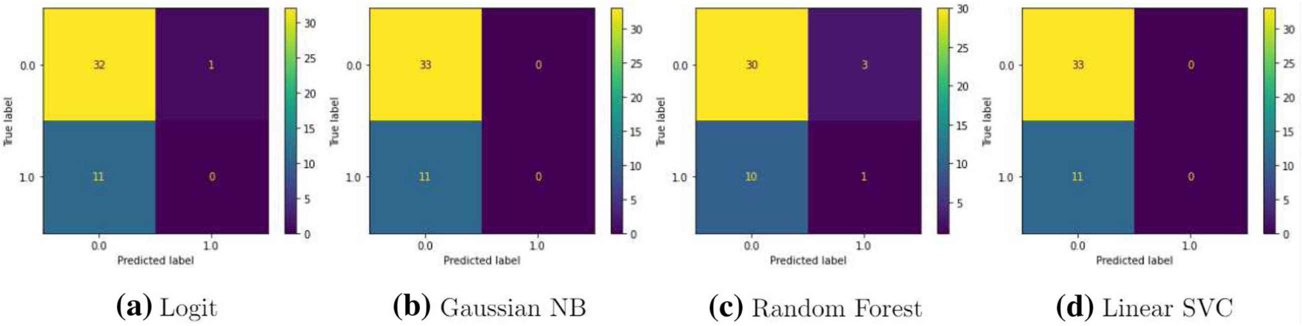
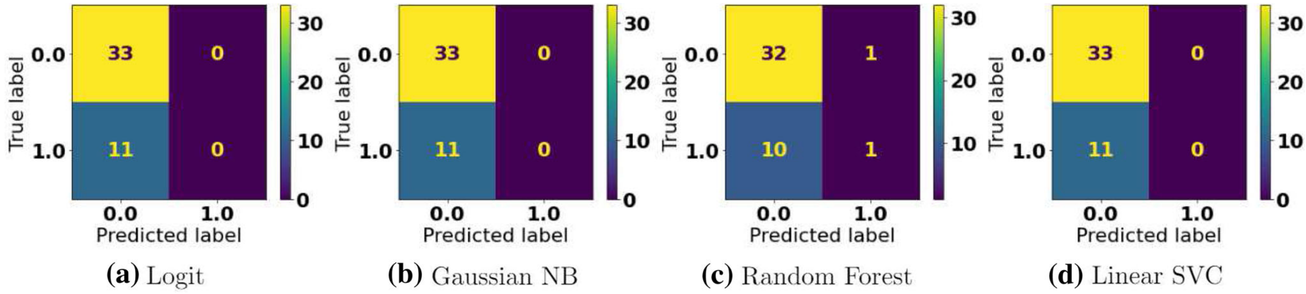


Fig. 8 Confusion matrices for the tested classifiers without PCA



(a) Logit **(b)** Gaussian NB **(c)** Random Forest **(d)** Linear SVC

Fig. 9 Confusion matrices for the tested classifiers with PCA

Table 2 Classifiers measures without and with PCA

PCA Applied	Classifier	Class	Precision	Recall	F1 Score
Not Applied	**Logistic Regression**	Stable	0.74	0.97	0.84
		Progressed	0.00	0.00	0.00
	Gaussian NB	Stable	0.75	1.00	0.86
		Progressed	0.00	0.00	0.00
	Random Forest	Stable	0.75	0.91	0.82
		Progressed	0.25	0.09	0.13
	Linear SVC	Stable	0.75	1.00	0.86
		Progressed	0.00	0.00	0.00
Applied	**Logistic Regression**	Stable	0.75	1.00	0.84
		Progressed	0.00	0.00	0.00
	Gaussian NB	Stable	0.75	1.00	0.86
		Progressed	0.00	0.00	0.00
	Random Forest	Stable	0.76	0.97	0.85
		Progressed	0.50	0.09	0.15
	Linear SVC	Stable	0.75	1.00	0.86
		Progressed	0.00	0.00	0.00

which resulted in a very good accuracy (75% on average), but very low precision and recall for the rarer class. Furthermore, while PCA had a very slightly positive effect on all classifiers, it was not able to solve the imbalance problem.

There exist several methods to solve the imbalance problem, but choosing the right one depends on the specific dataset. The first and probably most common approach is represented by applying a down sampling on the most common class, which can be generally obtained through a randomized selection of the original samples. However, this technique is generally applied when enough data are available, that is when losing some of the original samples would not translate into an important information loss. This is not our case: as we already have a very small dataset, down-sampling cannot be taken in consideration.

A second approach is represented by over sampling the rarer class, that is introducing new samples that reproduce the features of the existing ones from the least common class. This is a more applicable approach in our situation, as it would not reduce the dataset, but will expand it artificially. One huge drawback of this kind of approaches is that classifiers tend to overfit after its application, so it is always a good idea to validate the classifier after the training.

Here we have tested three different approaches for over sampling that are quite commonly used in the literature to tackle the imbalance problem with small datasets. Again, we have used the functions provided by Scikit-Learn to quickly apply them to our data. The tested approaches are:

- Random oversampling, with duplication of randomly selected samples from the rarer class. This is probably the simplest approach currently applied in the literature, and it proves to be quite effective in many cases.
- Synthetic minority oversampling technique (SMOTE) algorithm (Chawla et al. 2002), which synthesizes new data by applying a K-nearest neighbor classification on the available samples and then randomizing the new characteristics. SMOTE is probably the most used approach when it comes to the synthesis of new samples from existing ones, and several other algorithms have been built on it.

- Adaptive synthetic sampling (ADASYN) algorithm (He et al. 2008), which operates similarly to SMOTE but also uses the data and feature distributions to better synthesize the new samples.

After the application of the three algorithms, the dataset contains two balanced classes, with exactly 166 samples each. The Stable class has not been altered in the process, as it does not need to be re-sampled, while the progressed class is now composed of many replicated or synthesized elements.

Here we have applied both the resampling algorithms and the PCA dimensionality reduction, taking in consideration 5 principal components which retain the 99% of the original information. Among all of the examined classifiers, random forest is surely the best candidate: it attempted to predict the rarer progressed class despite the imbalance, indeed being the only one not having a 0.0 value for precision on the progressed class (see Table 3); it is generally less prone to overfitting, but here the problem needs to be addressed.

The graph reported in Fig. 10 shows that the accuracies of the classifiers change considerably after the application of the oversampling and that they are much more variable than before. As it results from the graphs, random forest has the highest accuracy among the considered classifiers, and its overfitting is almost null if we take in consideration the random oversampling. Other classifiers reach lower accuracies, but still show to be overfitting in the different configurations.

While the accuracy of the classifiers has dropped, with the only exception of random forest, Precision, Recall and F1 Score have no zero value for either class now, thanks to the re-balance .

This depends on the fact that the classifiers are not blindly selecting the classes anymore, as it is confirmed by the confusion matrices in Figs. 11, 12, 13 and by the measures reported in Table 3.

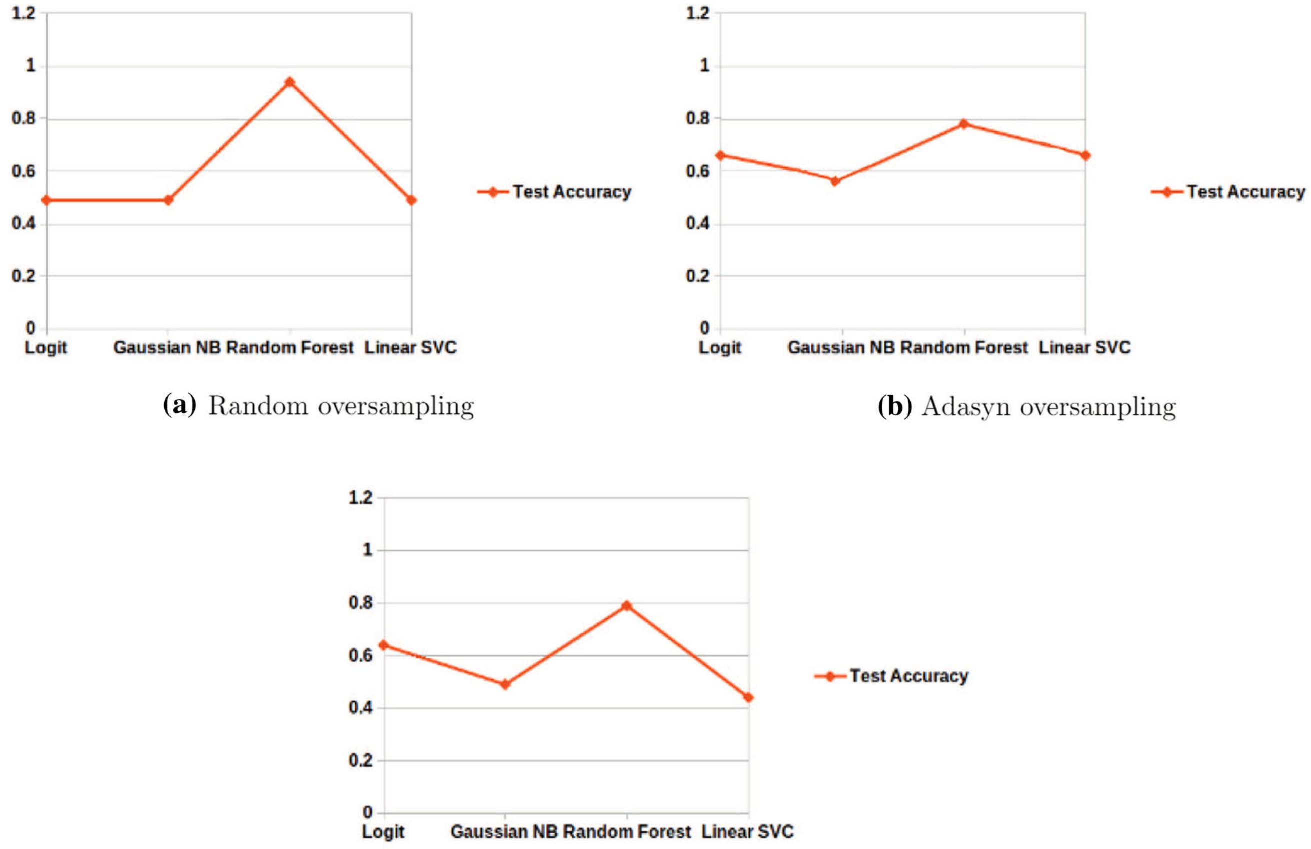
It is indeed evident by the fact that Precision and Recall is not zero anymore for the Progressed Class that the classifiers are working much better than before, with the random forest reaching very high accuracy values. In particular, random forest reaches an accuracy of 91% on the test set and of 99% on the training set, showing that overfitting is still occurring. The gap between accuracies is even wider with the SMOTE and ADASYN approaches, as also shown in Table 4.

6.2 Algorithm calibration

According to the preliminary measures taken and shown in Sect. 6.1, random forest is the best candidate for our clas-

Table 3 Classifiers measures after the oversampling

Technique	Classifier	Class	Precision	Recall	F1 Score
Random	**Logistic Regression**	**Stable**	0.52	0.40	0.45
		Progressed	0.47	0.59	0.53
	Gaussian NB	**Stable**	0.52	0.34	0.41
		Progressed	0.48	0.66	0.55
	Random Forest	**Stable**	0.97	0.86	0.91
		Progressed	0.86	0.97	0.91
	Linear SVC	**Stable**	0.52	0.40	0.45
		Progressed	0.47	0.59	0.53
ADASYN	**Logistic Regression**	**Stable**	0.80	0.53	0.63
		Progressed	0.59	0.84	0.69
	Gaussian NB	**Stable**	0.68	0.39	0.50
		Progressed	0.51	0.77	0.62
	Random Forest	**Stable**	0.86	0.63	0.73
		Progressed	0.66	0.87	0.75
	Linear SVC	**Stable**	0.80	0.53	0.63
		Progressed	0.59	0.84	0.69
SMOTE	**Logistic Regression**	**Stable**	0.67	0.59	0.62
		Progressed	0.62	0.70	0.66
	Gaussian NB	**Stable**	0.50	0.51	0.45
		Progressed	0.49	0.58	0.53
	Random Forest	**Stable**	0.74	0.76	0.75
		Progressed	0.75	0.73	0.74
	Linear SVC	**Stable**	0.46	0.53	0.49
		Progressed	0.43	0.36	0.39

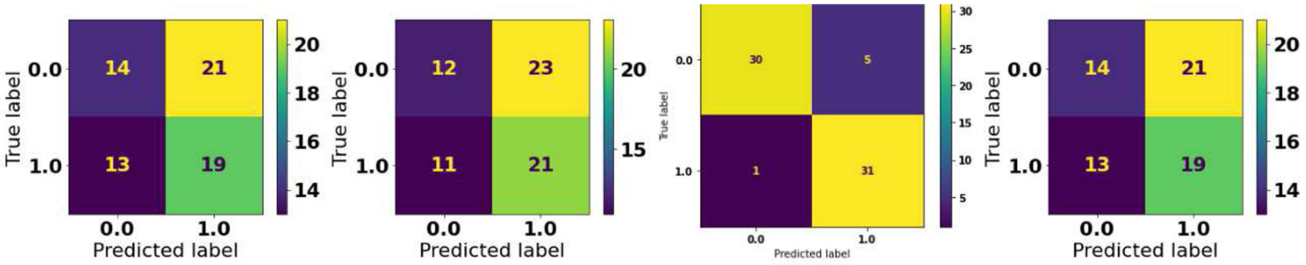


(a) Random oversampling

(b) Adasyn oversampling

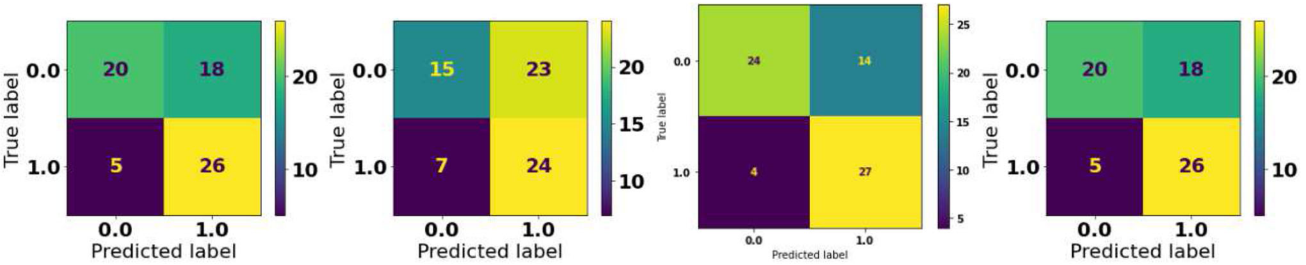
(c) SMOTE oversampling

Fig. 10 Accuracies for the tested classifiers after the oversampling



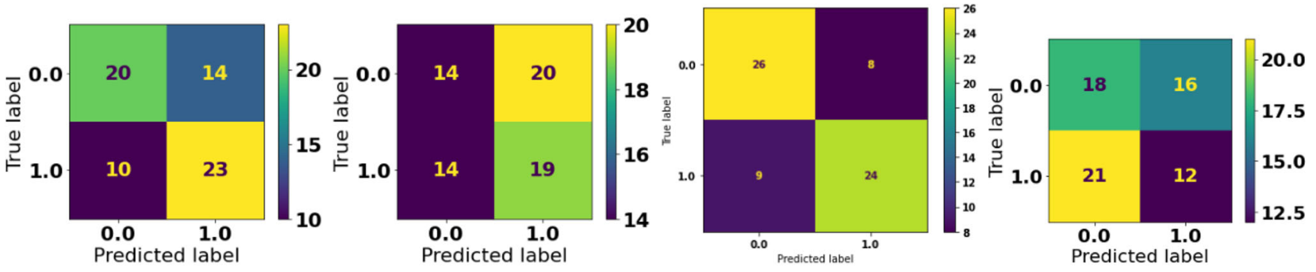
(a) Logit **(b)** Gaussian NB **(c)** Random Forest **(d)** Linear SVC

Fig. 11 Confusion matrices for the tested classifiers after the random oversampling



(a) Logit **(b)** Gaussian NB **(c)** Random Forest **(d)** Linear SVC

Fig. 12 Confusion matrices for the tested classifiers after the ADASYN oversampling



(a) Logit **(b)** Gaussian NB **(c)** Random Forest **(d)** Linear SVC

Fig. 13 Confusion matrices for the tested classifiers after the SMOTE oversampling

Technique	Classifier	Accuracy mean over 5 folds	Accuracy std over 5 folds
Random	Logistic Regression	0.50	0.028
	Gaussian NB	0.49	0.034
	Random Forest	0.85	0.048
	Linear SVC	0.50	0.030
ADASYN	Logistic Regression	0.57	0.043
	Gaussian NB	0.55	0.07
	Random Forest	0.73	0.038
	Linear SVC	0.57	0.043
SMOTE	Logistic Regression	0.55	0.038
	Gaussian NB	0.52	0.067
	Random Forest	0.71	0.011
	Linear SVC	0.60	0.042

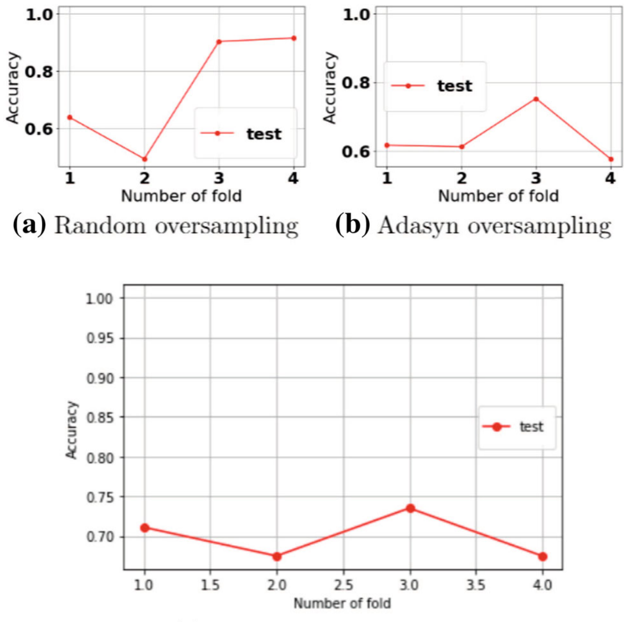
Table 4 Train and test accuracy for random forest with different rebalance approaches

	Trainaccuracy	*Testaccuracy*
textbfRandom upsample	0.99	0.91
ADASYN	0.99	0.74
SMOTE	1.00	0.75

sification problem. However, since it tends to overfit over the data, mostly due to the oversampling, we need to tune it appropriately.

First of all, we need to select one of the three oversampling techniques: in order to do so, we run the training of the random forest classifier by using the three expanded datasets, but we also use the **Ķ-fold** data selection algorithm in order to reduce the loss of data for testing and also to verify the effect of the oversampling over different folds of data. By applying the K-Folds technique, with $K = 4$, we obtain the results shown in Fig. 14.

The results show a clear overfitting over all folds, but the SMOTE approach seems to provide a more stable accuracy, which does not vary too much among the examined folds even if it is lower than in the random oversampling. Using the SMOTE algorithm seems to be a better choice for our purposes, since accuracy is not too dependent on the data separation.



(a) Random oversampling **(b)** Adasyn oversampling

(c) SMOTE oversampling

Fig. 14 Accuracy of random forest with K-Fold with different folds ($K = 4$)

7 Conclusion

The presented article describes the application of machine learning techniques to a very limited dataset, and all the necessary preprocessing that were run in order to obtain the acceptable results for the prediction of the future course of the multiple sclerosis disease in limited sets of patients. The objective is to provide additional instruments to practitioners that, in similar cases, can only apply statistical approaches that do not always provide the accurate results, due to the very small number of available samples. Starting from questionnaires submitted by patients in 2017, it was possible to extract a preliminary features vector to describe the input data, while information on the progress of the disease were obtained by monitoring the same patients in the following 3 years. With these data, different classifiers were initially trained, but they did not provide significant results, both due to the presence of redundant features in the dataset, and the unbalance of the examined classes. Due to the imbalance it was necessary to use oversampling techniques in order to increase the number of items of the least occurring class, after which the results became satisfactory. In particular, using the SMOTE oversampling technique and the random forest classifier, a precision of 79% is achieved with acceptable recall and accuracy values.

Despite the absence of a consistent dataset, our study successfully obtained good accuracy, precision and recall results, which are comparable and in line with the other cited similar studies, although our sample was smaller and the approach was based only on PROMs without clinical measures.

In the future, the methodology will be applied to larger datasets that will be collected through new clinical trials. Furthermore, data from patients monitored after a longer period, i.e., 6 years, will be taken in consideration. Also, as new data are collected, the trained algorithm will be tested on new patients to verify its degree of reliability and will be refined, in order to obtain better accuracy, precision and recall. Another aspect that will be investigated regards the application of Regression algorithms, or deep learning approaches such as LSTM. However, these will require larger datasets to be efficient enough, and will probably produce worse results in the first phases of the experiments, despite the application of dimensionality reduction and oversampling techniques.

Author Contributions All authors contributed equally to this work, and have worked on the design of the paper. DB, BdM and AE focused on the design and implementation of the machine learning framework. GT, SB and LL designed the problem and objectives of the paper, provided the dataset and evaluated the final results.

Funding Open access funding provided by Università degli Studi della Campania Luigi Vanvitelli within the CRUI-CARE Agreement.

Data availability The datasets generated and/or analyzed during the current study are not publicly available due to privacy reasons, related to the GDPR compliance.

Declarations

Conflict of interest The authors declare no conflict of interest.

Informed consent Informed consent was obtained from all subjects involved in the study.

References

Brichetto G, Monti Bragadin M, Fiorini S, Battaglia M, Konrad G, Ponzio M, Pedullà L, Verri A, Barla A, Tacchino A (2019) The hidden information in patient-reported outcomes and clinician-assessed outcomes: multiple sclerosis as a proof of concept of a machine learning approach. Neurol Sci. https://doi.org/10.1007/s10072-019-04093-x

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Confavreux C, Vukusic S (2006) Natural history of multiple sclerosis: a unifying concept. Brain 129(3):606–616. https://doi.org/10.1093/brain/awl007

Damasceno A, Pimentel-Silva LR, Damasceno BP, Cendes F (2020) Exploring the performance of outcome measures in MS for predicting cognitive and clinical progression in the following years. Multiple Scler Relat Disord 46:102513

Di Martino B, Colucci Cante L, DAngelo S, Esposito A, Graziano M, Marulli F, Lupi P, Cataldi A (2021) A big data pipeline and machine learning for a uniform semantic representation of structured data and documents from information systems of Italian Ministry of Justice. Int J Grid High Perform Comput (IJGHPC)

He H, Bai Y, Garcia EA, Li S (2008) Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pp 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Grothe M, Lotze M, Langner S, Dressel A (2016) The role of global and regional gray matter volume decrease in multiple sclerosis. J Neurol 263(6):1137–1145. https://doi.org/10.1007/s00415-016-8114-3

Mukaka MM (2012) A guide to appropriate use of correlation coefficient in medical research. Malawi Med J 24(3):69–71

Muthuraman M, Fleischer V, Kroth J, Ciolac D, Radetz A, Koirala N, Gonzalez-Escamilla G, Wiendl H, Meuth SG, Zipp F, Groppa S (2020) Covarying patterns of white matter lesions and cortical atrophy predict progression in early MS. Neurology—Neuroimmunology Neuroinflammation **7**(3)https://nn.neurology.org/content/7/3/e681.full.pdf. https://doi.org/10.1212/NXI.0000000000000681

Nasir IM, Khan MA, Yasmin M, Shah JH, Gabryel M, Scherer R, Damaševičius R (2020) Pearson correlation-based feature selection for document classification using balanced training. Sensors 20(23):6793

Rojas JI, Patrucco L, Alonso R, Garcea O, Deri N, Carnero Contentti E, Lopez PA, Pettinicchi JP, Caride A, Cristiano E (2021) Diagnostic uncertainty during the transition to secondary progressive multiple sclerosis: multicenter study in Argentina. Mult Scler J 27(4):579–584. https://doi.org/10.1177/1352458520924586

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Yaping Z, Changyin Z (2021) Gene feature selection method based on Relieff and Pearson correlation. In: 2021 3rd international conference on applied machine learning (ICAML), pp 15–19. https://doi.org/10.1109/ICAML54311.2021.00011

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.