Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 219 (2023) 1365–1372

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2022

Preliminary Results to Predict Tuberculosis Outcomes Applying Traditional and Automated Machine Learning Models

Ana Clara de Andrade Mioto[a,b1], Mariana Tavares Mozini[b], Renan Barbieri Segamarchi[b], Giovane Thomazini Soares[b], Pedro Emilio Andrade Martins[b], Victor Cassão[a], Luís Gustavo Barichello Ferrassini[c], Newton Shydeo Brandão Miyoshi[b], Domingos Alves[b] and Lariza Laura de Oliveira[b]

[a]*Bioengineering Postgraduate Program, University of São Paulo, São Carlos, Brazil*
[b]*Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil*
[c]*University Center Barão de Mauá, Ribeirão Preto, Brazil*

Abstract

Tuberculosis (TB) remains one of the most lethal infectious diseases in the world and, despite being preventable and curable, kills 4.500 people daily, according to the World Health Organization (WHO). Brazil, being a country heavily affected by TB, works to improve social intervention programs, since the decrease in the patients vulnerability seems to have a positive effect for the cure of TB. The Brazilian public health system records data on TB treatment that can guide actions and interventions. In this context, machine learning (ML) algorithms have been used successfully to analyze health and medicine (H&M) datasets. An emerging area of ML called Automated Machine Learning (Auto-ML) was tested in this analysis to predict the following TB results: good and bad outcomes. Our results indicate that it is possible to build reasonable ML models with the available data.

Keywords: Tuberculosis; Bad Outcomes; Machine Learning; Automate modeling;

[1]* Ana Clara de Andrade Mioto. Tel.: +55 16 99235-9934;
E-mail address: anaclara.mioto@usp.br

1877-0509
10.1016/j.procs.2023.01.424

1. Introduction

Tuberculosis (TB) is an infectious disease caused by the bacillus Mycobacterium tuberculosis that can be spread through the air from one person to another. It affects mainly the lungs, but when found in extrapulmonary form, TB can affect other organs or systems, especially on those with immune compromise. According to the World Health Organization (WHO), tuberculosis is the 13th leading cause of death worldwide and the second cause of death from a single infectious agent after COVID-19, being responsible for the death of 1.5 million people in 2020 including 214.000 people with HIV [1].

In 2014, the WHO launched The end of TB strategy aiming to end the global TB epidemic by setting goals to be met by 2035, which includes 95% reduction in the number of TB deaths and 90% reduction in TB incidence rate compared with 2015 [2]. Despite the progress made over the years with the strengthening of existing TB services [3] and the development of several researches, about a quarter of the global population is still infected with TB bacteria with most of these cases occurring in emerging or underdeveloped countries [4].

The relationship between social inequality and tuberculosis has been well established and became evident since the list of high burden countries comprises the ones with the largest social inequality indicators. Poverty, undernutrition, lack of access to the health system are underlying risk factors for TB. For this reason, Brazil has worked to improve social protection intervention programmes such as cash transfer programs, and universal eligibility for TB free treatment, which can have a positive effect towards TB cure [5].

In 2020, Brazil registered almost 70 000 new cases of TB, however the situation created by the new coronavirus has changed the epidemiological indicators such as the reduction in total TB notifications at the three levels of care and the reduction of molecular fast tests consumption. In addition, cure indicators amongst the new cases of pulmonary TB in 2019 reached 70,1% depending on the region and 12% of the new confirmed cases (two times higher than the recommended by WHO) abandoned treatment [6]. According to the recommendation of the Brazilian Ministry of Health, rates such as those mentioned previously indicate the need to increase the quality of treatment coverage, especially because patients who abandon treatment contribute to disease transmission [7,8].

Many computational solutions can be applied in the field of medicine to deal with healthcare data, where the intent is to facilitate and improve medical treatment, reduce costs and promote advances in medical research [9]. Through the application of Machine Learning algorithms, patterns from non-linear combinations like social demographic and clinical factors, obtained from the tuberculosis data collected across the state of São Paulo, can be used with the intention to prevent bad outcomes from occurring and to guide actions to be taken before and after their occurrence.

In this paper, we tested classic ML methods against an AutoML approach to detect TB outcomes, trying to evaluate and discuss how AutoML can be useful in the TB data context. The main objective of this analysis is to improve the prediction of some TB outcomes. As a result, we expect to be able to detect treatment abandonment before it happens and predict risks of death caused by TB, considering only the early information of the patient, collected in the beginning of the treatment.

2. Background

2.1. Machine Learning

Thanks to technological advancement, many more ways to treat and to understand data have been developed, creating the study field of data science. There was a necessity to go beyond the human brain, find relations invisible to our eyes, and with this will, the machine learning (ML). Nowadays, with this kind of application, it's possible to correlate patterns between variables that humans haven't imagined.

Using ML models helps a lot on the development of software, removing the urge to specify many rules for a simple decision, designing the whole system with the exact code for those specificities. With ML, the system can learn from itself, giving to the project, more liability, and more flexibility [9].

Going deeper into ML techniques, there are two ways to apply this technology: the supervised and the unsupervised learning. This first one gives its results with the system decision-making based on some examples entered on the input of the model, whose input contains many of the desired outputs, and the proper algorithm to analyze the data and provide the correct output. A great example for this method is the spam tag in the mailbox. This

is an algorithm of ML, in which the developers input some emails, and by the code, the machine learns which ones are spam, and which ones are not, based on the user's action, and on some previous content from the e-mail.

Unsupervised learning works in a different way. There are no previous outputs declared to the algorithm, it's known only the input data, but not the output, which the machine will return, by analyzing the inputs, and their relation. An example for this one is the process of segmenting groups of people, ordering by preferences, it's known only what people prefer, but how the algorithm is going to group it, only by its output [10].

2.2. Automated Machine Learning

Having this background in ML, it has been developed the Automated Machine Learning (AutoML), which is based in the theory of the traditional ML, but with some improvements, going further to the massive use of machines. AutoML gives the developers the opportunity to create many models, with no effort. The performance of ML algorithms was highly improved, with a process called Hyperparameter optimization (HPO), in which it is possible to adapt the same model to different kinds of situations. Even though some HPO processes may be very expensive, their benefits surely pay off, due to the facility for making comparisons between different methods of treating data, since they've received the same treatment [11].

Auto ML also provides stronger results, not in terms of liability, but in terms of being faster. By using the correct generation of the models, with some amount of time, the hyperparameters could be unimaginable. Called baselines, a group of similar algorithms can co-work, and cut a big time before creating the models.

An iconic example of this is Erin Ledell, a programmer that created a baseline during Kaggle's Hackathon, with only one line of code and 1h40min of pure training, she obtained an amazing position in the competition [12].

2.3. Tuberculosis

The actions envisaged by the End of TB Strategy are essential pillars to end TB across the globe, and range from patient centered TB care and prevention to research and supportive policies. All of these require universal use of digital health tools that can improve quality, effectiveness and efficiency of their efforts [13].

There is an emergency for platforms and research practice that deals with large amounts of data and sophisticated computational methods for scientists to collaborate. Research in eScience addresses all aspects of the research processes (data collection, simulations, modeling, creation of tools), it is a multidisciplinary work where computer science is able to help researchers to develop research in a faster and more effective manner [14].

In Brazil, the Research Program in eScience, created in 2013 by São Paulo State Research Support Foundation (FAPESP), is an example of initiative that works towards the goals set by WHO of innovative and unconventional approaches to scientific research [15]. In this context, this article is a research on eScience as it contributes to the expanse of technology and innovation to medicine and healthcare fields, where broader determinants of TB, like poverty and comorbidities, can be taken into consideration in the prediction of outcomes, helping not only the patients and hospital facilities but also to contain the dissemination of the disease.

3. Methodology

3.1. Dataset

For the development of this study, we used a sample of around 3000 patients, collected from the TBWeb health information system, between 2016 and 2019. The database, which has been increased, belongs to a study being conducted involving all registries and their different bad outcomes. Based on the need to test methods and tools that can help in this larger study, we separated this sample and applied all the stages of a knowledge discovery process, or rather, data science project, using the python programming language.

3.2. Preprocessing Step

The step after data collection involves all preprocessing, starting with the identification and treatment of inconsistencies. This dataset has 113 attributes and all sensitive data were anonymized. Inconsistencies and missing values were also removed and identified as the symbol. The values not adequate to what the column represents were analyzed each one in its context checking for consistency, replacing them with "Ignored", "No information" or "Others". This will not only clean the data, but organize it. Here we used the Pandas and Numpy packages [16,17].

Then we perform a check for duplicate values. For this study, we chose to remove the lines that contained most of the empty information and that were duplicated, as it did not allow the extraction of any knowledge there and affected the learning of the model. For better visualization and following good programming practices, we standardized the reading of the columns and the scale of the data, converting them to uppercase and objects to numeric (when necessary), highlighting that Python is case sensitive, so we can avoid future errors. To finish this stage, based on the literature [18,19], we selected the variables that make the most sense for our models and the objective we have with this research, resulting in 46 variables, the most important ones are listed in Table 1.

Table 1. Final attributes from TBWeb System.

Column Name (Portuguese)	Column Name (English)	Type of Attribute
RACA_COR	Ethnics	categorical
IDADE	Age, separated into periods	categorical
SEXO	Gender	categorical
GESTANTE	Pregnant	boolean
ESCOLARIDADE	Education, type in years of education	categorial
TIPO_OCUPACAO	Work occupation	categorial
BACILOSCOPIA_ESCARRO	Sputum smear	categorial
HISTOPATOLOGIA	Histopathology	categorial
HIV	HIV	boolean
AIDS	AIDS	boolean
DIABETES	Diabetes	boolean
ALCOOLISMO	Alcoholism	boolean
DOENCA_MENTAL	Mental disease	categorial
USO_DROGAS	Use of drugs reported	boolean
OUTRAS_DOENCAS_IMUNO	Immune disease	boolean
TABAGISMO	Smoking	boolean
RESISTENCIA	Resistant TB	boolean
SITUACAO_ATUAL	Outcome (good or bad)	categorial

3.3. ML applied to TB classification

After the preprocessing step, described, we selected the sample mentioned above. Of these, 70% are male and most of the cases don't have Aids or HIV as a comorbidity. Some other highlights observed during the exploration of the data are that 65% of patients have low education (between 4 and 11 years old) and 50% of all cases are young, aged between 20 and 39 years.

Regarding our target variable: outcome (in portuguese 'SITUACAO_ATUAL'), we performed a brief treatment to transform it into binary, following the WHO classification of what is a good and a bad outcome [20]. As a result, we obtain the following division, being 1 for good, composed with cure, in outpatient treatment and in inpatient treatment, and 0 for bad, containing abandonment, death no TB, diagnosis change, death TB, transfer to another state/country, resistance, change of scheme due to intolerance/toxicity, transfer, primary abandonment, absentee, other and no information.

Before releasing all the data into the models, we observed that the dataset has an imbalance between the target variable, that is, only 18% of the patients have bad outcomes. With this imbalance in mind, we decided to apply the machine learning models to the balanced and unbalanced dataset, obtaining the results shown in Tables 2 and 3, respectively. The balance was made with the help of Near Miss technique (based on under-sampling) [21], the remaining data was normalized.

During classification, we tested KNN (K nearest neighbors), and Random Forest from the Scikit-learn package [22] against the AutoML approach, from the PyCaret package [23]. The main metrics analyzed during the study of

the ML and AutoML models were accuracy, F1 score, precision, and the most important for the research, recall score, this last one because our targets are the outcomes for the patient, so using an evaluation based on sensibility will improve the results expected. After treating data, it was needed to filter some features, just leaving those with more interest for the machine learning to do its job. All models were trained considering a split of 70% for the training set and 30% for the testing set.

4. Results

4.1. ML applied to TB classification

It is interesting to note that the models had good results with unbalanced data, as seen in the F1-score of 0.91 and sensitivity of 0.98 in the Random Forest model. In relation to the balanced data, we obtained a relatively low accuracy in both models, with 0.69 for KNN and 0.78 for Random Forest, but they achieved good recall and F1-score, 0.81 and 0.82 for KNN, 0.83 and 0.79 for Random Forest, in that order.

Table 2. Results with balanced data.

Model	Accuracy	Precision	Recall	F1-score
K-Nears Neighbors (KNN)	0.69	0.65	0.81	0.72
Random Forest	0.78	0.75	0.83	0.79

Table 3. Results without balanced data.

Model	Accuracy	Precision	Recall	F1-score
K-Nears Neighbors (KNN)	0.79	0.85	0.89	0.87
Random Forest	0.84	0.85	0.98	0.91

4.2. ML applied to TB classification

Table 4 shows the results of the best classifiers found by PyCaret. The package allows automatic data balancing, by setting this option. So with that in mind, we also tested it with the balanced data afterwards, and the technique automatically chosen by PyCaret was SMOTE (over-sampling). We included both results, considering with and without the automatic data balanced as can be seen in Table 5. One can observe that best accuracy values were found for Random Forest (0.8423).

Table 4. Results with PyCaret - imbalanced.

Best Models	Accuracy	Precision	Recall	F1-score
Random Forest Classifier	0.8423	0.8470	0.9855	0.9095
Extra Trees Classifier	0.8328	0.8420	0.9796	0.9109

Table 5. Results with PyCaret - balanced.

Best Models	Accuracy	Precision	Recall	F1-score
Random Forest Classifier	0.8499	0.8564	0.9839	0.9157
Extra Trees Classifier	0.8437	0.8561	0.9753	0.9118
Light Gradient Boosting Machine	0.8552	0.8864	0.9465	0.9154

5. Discussion

It is important to highlight that was not possible to obtain a ”perfect” model for the prediction of bad outcomes, a possible reason for this is that the attributes used may not be predictive, i.e. the initial patient data recorded in the beginning of the treatment may not be enough to predict this bad outcome (The used attributes can be seen in the Subsection pre-processing), other important point is that we used a sample, not being a representative amount of data. It is also known that the causes of abandonment, for example, are complex and according to Santos et. al [24], they are directly connected with the service coordination and with how care professionals conduct their actions with families and patients, as well as how TB patients are followed-up throughout the treatment [24]. Another important question, pointed in [24], which is difficult to measure, is the lack of comprehensiveness of the health care process by the patient caused by a medication-focused treatment which sometimes ignores important aspects as the social economic ones, known to be important [25]. This impact in the perception of his own disease and treatment difficulties the adherence to the treatment.

Despite the difficulties in obtaining a ”perfect” model, our results were very promising regarding the use of AutoML. This can be seen when comparing the results obtained with regard to balancing, in which the recall was much better in AutoML, thinking that both presented Random Forest with the best results (classical : 0.83 < AutoML : 0.98).

It is precisely this balance that we must pay attention to when working with real data, which in most cases is unbalanced. That’s why we chose to present the results for both cases, which raised another question: why are unbalanced results better than balanced ones? A possible explanation is the number of examples of outcomes available for training the model, that is, the dataset became less representative with the class of our interest, being much higher in the case of good outcomes for classic machine learning, remembering that we chose to use the Near Miss technique, using as much real data as possible, so that we do not have a high amount of synthetic data being created, such as occurred in AutoML that chose to use the SMOTE technique.

Tables 4 and 5 show the best models found by PyCaret, it can be seen that in both, there were selected models that are ensembles and showed high results for the recall. An important question raised when analyzing these generated models is the interpretability of them. Recently, several researchers have questioned the validity of models with high values of predictive accuracy but poorly interpretable. In the medical domain where interpretability is crucial, losing accuracy can be acceptable for the benefit of an interpretable model, as pointed out by [26,27]. Considering interpretability, models can be classified as white box (interpretable), black box (non-interpretable) and gray box (partly interpretable). Considering separately the classifiers composing the final models selected, most of them can be considered as gray boxes (Random Forest, extra trees) and some of them black boxes. In addition, the very fact that the final model is an ensemble would make interpretation difficult.

6. Conclusion

In this preliminary analysis, we applied traditional and automated machine learning to predict two tuberculosis outcomes: good and bad, using the treatment and notification data, collected in the beginning of the discovery case. Our results suggested that it is possible to predict the final situation of TB treatment satisfactorily (higher F1-score values equal or major than 0.70).

In the best of our knowledge, we could not find any study that used automated machine learning techniques to predict outcomes of Tuberculosis. Researchers have successfully used only classical machine techniques to study Tuberculosis, for example, by analyzing chest radiography using deep learning techniques.

The main strengths of this analysis are the exploration and comparison of AutoML into medical context, particularly in the study of tuberculosis. A very important question raised was the interpretability of the models found. Since we did not limit the number of classifiers, the final model produced was an ensemble composed of basically gray and black box models. One way to handle this situation is to change this number to one, allowing a single model to be selected. Thus, as PyCaret also has interpretable models in its composition, we could choose from the best selected models those that were white boxes, analyzing trade-off between accuracy and interpretability, as proposed by.

As a conclusion, improvements need to be made to the model selection to increase sensitivity and prioritize the interpretability of generated models, to help the recognition and attendance of patients most vulnerable to

abandonment and death (these mainly among the bad outcomes), contributing to more accurate decision making and with more satisfactory results in the control of tuberculosis in health.

As future work, we propose to normalize the outcomes of study according to globally accepted ones, with the World Health Organization information and independently analyze each of the possible outcomes and we will analyze with more detail about the differences between balanced and imbalanced data. Another point to consider is that our set of attributes is relatively limited to local demographic information and some clinical information obtained at the beginning of the case. Given the complexity of factors associated with the bad outcomes of treatment, we hope that with a more elaborate set of information, including socioeconomic aspects, we will have better models.

Acknowledgements

This work in part was supported by the São Paulo Research Foundation (FAPESP) - grant number 2020/01975-9 and 2021/01961-0, coordinated by author DA.

References

[1] WHO: Tuberculosis,https://www.who.int/health-topics/tuberculosis. Last accessed 18 Jan 2022

[2] Global Tuberculosis Programme: The end TB strategy. World Health Organization, Geneva (2015)

[3] Pelissari, D. M., et al: Identifying socioeconomic, epidemiological and operational scenarios for tuberculosis control in Brazil: an ecological study. BJM Open, 1–10 (2018)

[4] World Health Organization and others: Global tuberculosis report 2021 (2021)

[5] Oliosi, J. G. N., et al: Effect of the Bolsa Familia Programme on the outcome of tuberculosis treatment: a prospective cohort study. The Lancet, 219–226 (2019)

[6] Secretaria de Vigilância em Saúde: Boletim Epidemiológico de Tuberculose. 1nd edn. Ministério da Saúde, Brásília (2019)

[7] Kritsk, A., et al: Tuberculosis: renewed challenge in Brazil. Sociedade Brasileira de Medicina Tropical, 2–6 (2018)

[8] Secretaria de Vigil ̂ancia em Saúde: Manual de Recomendações para o Controle da Tuberculose no Brasil. 2nd edn. Ministério da Saúde, Brasília (2019)

[9] Müller, Andreas C., Guido, Sarah.: Introduction to Machine Learning with Python: A Guide for Data Scientists. 1st edn. Publisher, O'Reilly Media, Inc., United States of America (2017).

[10] Pinheiro, Carlos A. R. and Patetta, Mike. 2021. Introduction to Statistical and Machine Learning Methods for Data Science. Cary, NC: SAS Institute Inc.

[11] Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)

[12] Medium, DataHackers, https://medium.com/data-hackers/automated-machinelearning-automl-parte-i-1d3219d57d31. Last accessed 22 Jan 2022

[13] World Health Organization: Digital health for the End TB strategy: progress since 2015 and future perspectives. World Health Organization, Geneva (2017)

[14] Appel, A. L.: A e-Science e as atuais práticas de pesquisa científica. Universidade Federal do Rio de Janeiro (5), 1–88 (2014)

[15] Programa FAPESP de Pesquisa em eScience e Data Science, https://fapesp.br/escience/. Last accessed 17 Feb 2022

[16] Pandas Python Package, https://doi.org/10.5281/zenodo.3509134. Last accessed 17 Feb 2022

[17] Numpy Python Package, https://www.nature.com/articles/s41586-020-2649-2. Last accessed 17 Feb 2022

[18] Hokino Yamaguti, V. et. al. Development of CART model for prediction of tuberculosis treatment loss to follow up in the state of Sãoo Paulo, Brazil: A case–control study. International Journal of Medical Informatics 141,(2020)

[19] Hokino Yamaguti, V. et. al. Charlson Comorbidities Index importance evaluation as a predictor to tuberculosis treatments outcome in the state of São Paulo, Brazil. Procedia Computer Science 138. pp. 258-263. (2018)

[20] WHO: Patient Safety, www.who.int/patientsafety/taxonomy/icps full report.pdf. Last accessed 17 Feb 2022

[21] Imbalanced Learn, https://jmlr.org/papers/v18/16-365.html. Last accessed 17 Feb 2022

[22] Scikit Learn Python Package, https://arxiv.org/abs/1309.0238. Last accessed 17 Feb 2022

[23] PyCaret Python Package, https://www.pycaret.org. Last accessed 17 Feb 2022

[24] Santos Alves, R., Mendes Jorge de Souza, K., Andrade Virgínio de Oliveira, A., Fredemir Palha, P., de Almeida Nogueira, J., et al.: Tuberculosis treatment abandonment and comprehensive health care to patients in the family healthcare strategy. Texto Contexto Enfermagem 21(3), (2012)

[25] Oliosi, J.G.N., Reis-Santos, B., Locatelli, R.L., Sales, C.M.M., da Silva Filho, W.G., da Silva, K.C., Sanchez, M.N., de Andrade, K.V.F., de Araújo, G.S., Shete, P.B., et al.: Effect of the bolsa familia programme on the outcome of tuberculosis treatment: a prospective cohort study. The Lancet Global Health 7(2), e219–e226 (2019)

[26] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1721–1730. ACM (2015)

[27] Freitas, A.A.: Automated machine learning for studying the trade-off between predictive accuracy and interpretability. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 48–66. Springer (2019)