Alma Mater Studiorum - University of Bologna

Machine learning classification of patient response in front of oncological therapy cycles

Zanatto Paolo

Master's degree in Digital Transformation Management
Machine learning - Prof. Guido Borghi
A.Y. 2022/2023

July 24, 2023

Abstract

The goal of this project is to predict the response in a defined period of time of a patient with a specific condition in front of a drug administration in therapy cycles. The analysis is focused on the response in 2 time period: 1 day and 3 days before the scheduled drug administration

1 Introduction

This project is done in collaboration with the Healthcare Administration department of IRST (Istituto Romagnolo per lo Studio dei Tumori "Dino Amadori"), that is a center of excellence entirely dedicated to care, research, and training in the field of oncology.

To understand the analysis is important to describe briefly the context with some information about cancer therapies: they work by cycles and each administration has an impact on the blood values of the patient, who is not able anymore to receive the drug for a period of time. This period of time depends by many factors and there are no rules to define it precisely; so before any administration it is planned to do a blood test in the day or days before the therapy execution, in order to check the patient condition by verifying the level of certain blood values This project aims to build a Machine learning algorithm that allows to understand if a patient in a certain condition, described by some information about him and therapy, is able to respect the schedule of the blood tests required to administer the cycles.

The IRST operations about oncological therapies work 5 days per week (instead the "traditional hospital" works every day, obviously), and the therapy administration works in different ways when it is planned on the last 4 working days (Tuesday, Wednesday, Thursday and Friday) where the day before the hospital "works"; compared to the first day of the week, where the blood test to authorize the therapy is done 3 days before (on the previous Friday). Given that, the dataset is full of therapies where the blood test has been executed 3 days before, and it is a really precious information. One of the goal of the Healthcare Administration department is to anticipate the blood test of one or 2 days in order to plan the drug production and optimize the warehouse management. With this goal, the percentage of success for the cases where the blood test has been executed 3 days before is a crucial information to understand if it is feasible to anticipate all the blood tests 3 days before instead of 1. The second goal of this project is to test if there is a percentage of failures (of the blood test) for the cases described is higher or lower compared to all the therapies

The goal is to build a classifier that is able to recognize if a row belongs to one of the 2 classes:

- Patient and condition where the blood test will authorize the therapy execution in the scheduled time (in the code "ready for the therapy").
- Patient and condition where the blood test will not authorize the therapy execution in the scheduled time (in the code "not ready for the therapy").

Considering the goal, the study is going to focus on the cases where the blood test is done 1 day before (the therapy scheduled from Tuesday to Friday) and 3 days before (the therapy scheduled on Monday). In detail, the project aims to verify the following hypothesis: "if the blood test is done 3 days before instead of 1, is the same the probability to authorize to execution of the therapy?".

1.1 Dataset

Each row of the dataset represents a therapy with some attributes about the patient and the blood test required to make the therapy. The data are extracted from the databases of IRST from the start of 2022 to the end of June 2023. The data collection system in IRST is based on web applications that allow to doctors, chemists and any operators of the hospital to operate.

The dataset is organized with the following attributes:

- IDPaziente: ID code for each patient
- genere: patient gender
- classeeta: age range
- IDTerapia: ID code for the therapy of a specific patient (eg. same therapies to different patients have different ID)
- CodiceSedeTumore: code to identify the area where the cancer is
- PrincipiAttivi: ID code for the active ingredient used in the therapy
- PrincipiAttiviDescrizione: active ingredients name
- NumeroSomministrazione: an incremental number about the step in the cycle of therapy
- DataSomministrazioneProgrammata: it should the scheduled date to execute the therapy cycle, but surely it is has been a modified before or during extraction (so it will not be used)
- **DataSomministrazione**: the actual date when the therapy cycle has been executed.
- DataSomministrazioneProgrammataPrecedente: it should be the scheduled date for the previous therapy cycle, but surely it is has been a modified before or during extraction (so it will not be used)
- **DataSomministrazionePrecedente**: the actual date of the previous therapy cycle.
- **GiornoDaTerapiaPrecedente**: the actual time passed between this therapy cycle and the previous one (DataSomministrazione – DataSomministrazionePrecedente).
- **GiorniTeoriciDaTerapiaPrecedente**: the expected time required to execute this therapy cycle and the previous one (real scheduled data for the current therapy cycle - DataSomministrazionePrecedente).
- PrimoPrelievo: the date of the first blood test
- **UltimoPrelievo**: the date of the last blood test.
- **NumeroPrelievi**: the count of the blood tests executed in order to authorize the current therapy step (if bigger than 1 it means that the first is failed).

2 Related work

As said in the introduction, every administration has an impact on the blood values of the patient, who is not able anymore to receive the drug for a period of time. This period of time depends by many factors and there are no rules to define it precisely, but there are some cases that has tried to use a ML algorithm to predict this information.

2.1 Prediction of recovery based on frequent monitoring

The first paper considered is the "Model-based prediction of myelosuppression and recovery based on frequent neutrophil monitoring" (2017) of Ida Netterberg1 et al. . This study was focused on monitoring the absolute neutrophil counts (ANC) during myelosuppressive chemotherapy, that applied with model-based predictions can improve therapy management. The model built is based on "Model of chemotherapy-induced myelosuppression with parameter consistency across drugs" (2002) Friberg LE, Henningsson A, Maas H et al; where it worked initially with simulated data coming from individual parameters and then it is improved by the data collected (by the monitoring activities). Jointly to the model, to make the predictions about individual parameters and individual ANC profiles, it used Bayesian classifiers. As a conclusion of this work it is possible to infer that the daily monitoring of the ANC, together with model-based predictions, could improve anticancer drug treatment by identifying patients at risk for the next drug administration and predicting when the next cycle could be initiated. About the impact of delaying the administration of a therapy, the paper considered how it can affect the patient mentally and cause unnecessary stress, with any other inconvenience for the clinic, administratively and/or financially, so forecasting a few days in advance the ANC can help to individualize the day of dose administration instead of postponing for a longer period.

2.2 Deep-learning-based prediction of absolute neutrophil count

Another study comes in 2022 from Hyunwoo Choo, Su Young Yoo et al. in "Deep-learning-based personalized prediction of absolute neutrophil count recovery and comparison with clinicians for validation". Before this work, predicting the recovery from myelosuppression mainly relied on pharmacokinetic-pharmacodynamic (PK-PD) models and mathematical modelling by Friberg et al. (quoted also in the other paper); but this method requires some empirically derived methods which limits its application in most clinical cases. To face this issue, they trained a deep learning model and selected the best performing model for predicting personalized ANC recovery after high-dose chemotherapy in pediatric patients, using data from 525 patients with solid tumors to predict the day when patients recover from severe neutropenia after high-dose chemotherapy. Overall, there were 51 variables, including blood cell count with differential and chemistry panels, name of the chemotherapeutic agents, and many others. At the end of the work, the model has been able to predict the exact end day of neutropenia (the condition that blocks the administration of therapy) for 24.24% of the cases. Then it shows errors within 1 day and 2 days of 76.76% and 94.94%, respectively.

2.3 My work

These are some of the work done in the last years related to the topic, especially they focus on predicting the recovery time required for each patient considering many data gathered over time on patients. Both of them had access to large amount of data and the collection of them has been studied in detail, Instead, my work is a bit different because it starts with different purposes: the data are generic and goal is not to understand the time required to recover the "blood value"s, but it aims to distinguish which are the conditions that allow to make the blood test one day or 3 days before the scheduled time to make the sequent therapy cycle. The data available are less and the algorithm has a more rough approach to the real events that occur inside the patients body, but exactly for this reason it could a better investment because the costs to collect data are zero. Furthermore, this approach can be improved in the future, especially when there is a better condition to collect additional data (compared to the existing one where IRST is facing an important transition and the recent environmental disaster had slow it down).

3 Proposed method

To reach the goal of this project, it has been used Colab to build a Python Notebook able to cover all the steps required to extract, prepare and execute the prediction to answer to the expected questions . The prediction has been done through 4 different Machine learning algorithms set with the parameters that fit as most with the data: they have been trained, validated, tested and finally compared to see which has been the best to create a classifier.

Instead, to answer to point about the blood test executed with 1 o 3 days before, it has been considered to run the code multiple times in order to verify 3 different conditions:

- Blood test done 1 day before (Tuesday-Friday theapies)
- Blood test done 3 days before (Monday theapies)
- Blood test done at most 3 days before (All the theapies)

After the execution of these 3, it has been possible to compare the percentage of "therapies with only one blood test executed" over all the cases.

The algorithm built is exhibited by the phases that compose it:

1. Data collection and acquisition.
   (a) The data are extracted from the databases of IRST from the start of 2022 to the end of June 2023. The data collection system in IRST is based on web applications that allow to doctors, chemists and nurses to operate. These data input works by "web page input objects" (buttons, select window, . . .) or by textbox. The details about the collection of the data used are not accessible, so it is assumed that can be many errors in the data.
   (b) For what concern the data acquisition, the extraction and the usage of data required the application of some privacy measures in order to make the usable: the age has been substituted with a "age-range", the "fiscal code" has been removed (by keeping a patient ID that don't involve any precise information), and the name of the cancer has been replaced with a code translatable only by IRST.
   (c) Talking about the code, the data has been uploaded in a structured way by means of a dictionary, where the field "PrincipiAttivi" is built as a list of strings.
2. Pre-processing. The adaptation of data has been one of the crucial phases of this project since the data received were not clean and usable by ML algorithm. It has been fundamental to understand the meaning of each field and to detect the rows that have to be corrected/deleted. These are the verifications carried out:
   (a) Duplication of values.
   The data extracted comes from a DB that, as known, guarantee the uniqueness of data. But my data are extracted by a transaction that I don't managed personally and that I cannot trust at all. At the end of this execution, the data appeared to be unique.
   (b) Missing values.
   There are many missing values in the fields "PrimoPrelievo" and "UltimoPrelievo". To solve this issue I considered the option to fill "UltimoPrelievo" by inserting the "DataSomministrazione"-1 for each case where the "GiornoDaTerapiaPrecedente" is equal to "GiorniTeoriciDaTerapiaPrecedente" (so the schedule has been respected) but this option was assuming that all these therapy executions

were scheduled on the last 4 days of the week. This is not a feasible assumption when you are interested to compare the results in percentage of blood tests done 1 day before and the ones done 3 days before, as this project aims to do. So, the rows where the "UltimoPrelievo" (and jointly "PrimoPrelievo") are missing has been deleted.

(c) Noise.
The collection phase can be one of the source of error, especially when this process is not clear. It has been applied a function to "clean" the data from the values that was clearly wrong.

(d) Creation of the attributes for the class.
The fundamental element in classification is obviously the class. In order to allow to algorithms to work it has been necessary to create a new attribute to detect the class in which each row belongs. It has been done by the use of a function that has verified the necessary conditions to assign a therapy in the classes of "Patient and condition where the blood test had not authorized the therapy execution in the scheduled time":

- If "GiornoDaTerapiaPrecedente" is higher than "GiorniTeoriciDaTerapiaPrecedente
- If "NumeroPrelievi" is higher than 1

When one of these two (at least) has been verified, the row is classified with 0, otherwise 1

(e) Creation of attributes for the days passed between blood test and therapy execution.
Still here, to make the analysis required it has been necessary to add an attribute that shows how many days have passed between the blood test and therapy administration. From this field are emerged that many rows have more days than the one considered by the problem (at most 3) and the causes of them can be predicted but they are not clear. Tt is possible to distunguish 2 cases:

- Distance blood test-therapy near to 3 (4 or 5): the blood test was probably done on friday (for monday) and the therapy execution has been postponed for other reasons (unknown)
- Distance blood test-therapy far from 3 (»3): data has been wrongly collected. Eg. the update of the date for the last blood test has not been done

These are some assumptions but there is no an easy way to verify them, so they cannot bring value to the algorithm and to the project. For these uncertainties, all these cases have been deleted.

(f) Management of the attribute "PrincipiAttivi" (containing arrays).
This attribute contains an array of strings for each active ingredients, but in this format does not suit with the algorithm because they are not able to manage different to numerical ones. To solve this issue, it has been created an attribute for each unique value of "PrincipiAttivi" (88 new attributes) containing 1 where that value (represented by the attribute) is present in the array and 0 where is not.

(g) Management of string attributes.
For the same reason of before, the attribute about "genere" and 'classeeta' have been translated into numerical information translated in this way: 'genere': 'F': 0, 'M': 1 'classeeta': '19-30': 0, '31-50': 1, '51-70': 2, '71-90': 3

(h) Elimination of attributes that cannot affect the classification.
The attributes of the dataset are many but most of them are just useful to extract the information useful for the project, that are:

- the class
- the time passed between blood test and therapy administration

Given that, the following attributes have been deleted:
'IDPaziente'
'IDTerapia'
'PrincipiAttivi'
'PrincipiAttiviDescrizione'
'NumeroSomministrazione'
'DataSomministrazioneProgrammata'
'DataSomministrazione'
'DataSomministrazioneProgrammataPrecedente'
'DataSomministrazionePrecedente'
'GiornoDaTerapiaPrecedente'
'GiorniTeoriciDaTerapiaPrecedente'
'PrimoPrelievo'
'UltimoPrelievo'
'NumeroPrelievi'

3. Modelling
As anticipated, to classify the objects has been considered to use some ML algorithms. The choice of do not consider any Deep learning approach is cause by the higher interpretability of ML outcomes, to ease the optimization of the algorithm, and I considered that the task is really complex (so potentially good for DL) but the attributes and information available are not sufficient to build a DL classifier. Coming back to the ML algorithms, the classification has been done by means of:

- Decision tree

- Support Vector Machines
- Ensemble methods
  - Random forest
  - Adaboost

The Decision tree has been considered because it is able to provided a graphical and clear description of the classification done and it appeared extremely useful when the model have to be discussed with "domain experts" or compare with other results. For this reason in the code it is printed a synthetic representation of the classification tree generated. Instead, the ensemble methods have been considered because the combinations of classifiers could strongly improve performance (compared to a single decision tree). Random forest has been considered because it should be able to build more stable classifier in front of stable training set; instead Adaboost would be able to detect better the cases "difficult" to be classified, and this dataset would be full of them considering the short availability of information (attributes) related to the data.
Each of the classifier worked with the parameters that fits as most with the data (eg. the minimum number of elements to create a leaf in Random forest trees) and choice of them has been able to classify the 10

4. The conclusion of the code is focused on comparing the errors done in the initial condition, where the therapies cycles are scheduled following the standard rules used over years (so the situation of the data and the current situation), with the errors that can be done if the therapies would be scheduled by means of the Ml algorithms studied (a possible future condition).
   Moreover it is used the confusion matrix to determine the quantity of misclassified elements and have the first support to determine the costs (and saved costs of using this approach).

4 Results

The results of the methods used are shown below.

1. Results only 1 day before
   Percentage of correct classified in the initial condition (just after deleting noise, missing values and duplicated values): 47.28%
   Percentage of correct classification of the best ML algorithm (SVC (C=20, gamma=0.1)): 65.01%

	Actually +	Actually -
Predicted +	1105	421
Predicted -	670	845

2. Results only 3 days before
   Percentage of correct classified in the initial condition (just after deleting noise, missing values and duplicated values): 45.17%
   Percentage of correct classification of the best ML algorithm (SVC(C=20, gamma=0.1)): 65.83%

	Actually +	Actually -
Predicted +	273	144
Predicted -	125	134

3. Results with blood tests <= 3 days before
   Percentage of correct classified in the initial condition (just after deleting noise, missing values and duplicated values): 46.38%
   Percentage of correct classification of the best ML algorithm (RandomForestClassifier(min$_samples_leaf = 4, n_estimators = 200)) : 65.54\%$

	Actually +	Actually -
Predicted +	1521	650
Predicted -	828	992

Looking at the confusion matrix it is easy to notice that in the cases of 1 days before and all cases (first and third tables) the quantity of "false-negative cases" (the ones predicted as negative but actually positive) are really high. It means that there was more ready people than expected, so more people is going to wait unnecessarily to make therapy. Also the "false-positive cases" are high, especially in the second graph. These cases are the ones that generate inconvenient for patients and also useless costs. Considering the costs question, probably the hospital would focus on the "false-positive cases", but before make this decision it is relevant to consider also how to reduce the "false negatives"; do not making it is would not be completely ethical.

The precision reached in this algorithm is not comparable with the ones of the papers analysed previously, but it is relevant to keep in mind the investment done to make this work and, mostly, the real necessary information for IRST. The approach of this work compared to other is rough, but maybe it has the necessary information for IRST to make the necessary improvements. So, given that, the precision reached must be related to the investment done.

5 Conclusion

Given the results it is possible to make some considerations.
About the comparison of 1 day and 3 days before blood tests it is easy to notice that the errors done in the past are really similar between 2 cases, with only 2.1% of difference. It is a good message because it means that the time scheduled to make therapy cycles are big enough to

have similar results in 2 days of difference. On the other hand, we can infer that the causes of blood test failures are not related strongly to the time, but there are factors much more relevant, where the future works should focus on.

About the implementation of this algorithm to improve the scheduling of blood tests and therapies administration. The data shows a clear improvement of around 20% (4000 cases, considering this data) and for each of them it could be possible to consider the cost savings in terms of:

- Material costs used to make the blood test
- Salary of nurses and operators involved in the service provision
- The inconvenience to the patient, that is undoubtedly hard to measure but it has a strong weight considering the role of hospital (where the contact and interactions problems are many)
- The reduction of pollution for the transport
- The improvements in warehouse management. The use of this prediction have an impact also on the product management for the increase of stability about the consumption of products (eg. shorter stock, less unexpected orders, . . .).
- An improvement in drug production. Being able to detect patient that can be ready for the therapy administration some days before, it allows to organize the production of therapies in order to maximize the usage of robot or spread the work to do in manual production of therapies (by chemists) on more relaxed time and avoid to have overwhelming ritms.

All these considerations requires to be deepen with the Healthcare Administration department and studied with the domain experts (eg. chemists and responsible of warehouse operations). So it is not easy to evaluate the real benefits on the operations of the hospital, but out of the real benefits, the costs to carry out this algorithm are low and it should be necessary only to find a way to integrate this one in the existing web application used to schedule blood test and therapies administration.

Even before, this algorithm could be improved by extracting more data (before 2022) or applying other techniques to the existing data. For instance it could be possible to improve the efficacy of predictions by grouping the active ingredients considering their impact on human body, so creating a hierarchy of groups and defining the best grouping solutions for maximize the classification. The logic would be: reducing the classes to improve the ability of classifier to recognize pattern. It could be an opportunity but it requires domain experts.