International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 12 (2023) pp. 176-188
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Received: 15 Jan. 2023; Accepted: 6 May, 2023; Published: 9 June, 2023

Data-Driven Prediction of Complications Risks in Cancer Patients: Machine Learning based approach

Imen Boudali [1,2]

[1] SERCOM Laboratory, University of Carthage,
Carthage 1054, Tunis, Tunisia
[2] National Engineering School of Tunis, University Tunis ElManar, Tunis Tunisia
imen.boudali@enit.utm.tn

***Abstract*: Cancer chemotherapy involves drugs that interfere with cellular functioning and lead to cell destruction. These cytotoxic drugs have narrow therapeutic index and in most of cases, their potential side effects concern directly and significantly non tumor cells. These adverse effects may be apparent in different forms of symptoms such as headache, nausea, breathing difficulty, tiredness, etc. In real cases, medical staff is facing difficulties to identify patients' state due to a lack of medical data. In order to limit chemotherapy related side effects and to support the medical staff in the clinical decision process, effective toxicity prediction and assessment structure are crucial. In this paper, we propose to assist treating physicians by predicting the toxicity level of each patient after each chemotherapy session. Thus, they early decide which drug adjustment is required and then prevent any further complication. Our support approach is based on machine learning techniques and relies on predefined toxicity levels for predicting chemotherapy complications. Multi-classification methods are considered and trained on real medical data that were collected during the treatment phase of cancer patients in Tunisia. An assessment of the proposed approach is performed through an experimental study to show the effectiveness and the performance of learning methods.**

Keywords: Clinical decision support, cancer patients, complication risks, data-driven prediction, supervised learning, multi-classification models.

I. Introduction

Nowadays, cancer diseases are the most leading causes of mortality worldwide. In Tunisia, a recent study on cancer mortality shows that the mortality rate is steadily increasing and tumors are the second largest leading cause of death after cardiovascular diseases [1]. Thus, cancer diseases are the most significant weakness to increasing life expectancy in all countries [2], [3]. Therefore, cancer research is an evolving field dedicated to saving lives. Early research focused on identifying cancer types before causing symptoms. Other researches focused on emerging new strategies for early prediction of treatment outcomes. Afterwards, with the emergence of information and communication technologies (ICT), collection of medical data have been provided for research community [4]. Cancer research concerns cancer biology, genomics, origins and causes, detection and diagnosis. In spite of inherent potentialities of ongoing efforts in this field, some issues are still challenging, especially therapeutic complication issues.

Generally, treating physicians needs a continuous control of the treatment progress for each patient in order to track side effects of their prescriptions during the chemotherapy phase. This tracking would optimize the intervention time to avoid any kind of complications [2].

According to medical studies, side effects during chemotherapy mostly depend on the following factors: type of drugs, the prescribed dose, the administration mode and the overall state of health. These effects occur despite the use of preventive medications [6].

In this work, our purpose is to support the medical staff during the therapy phase. Therefore, we initially focus on the study and analysis of the most common side effects for several types of frequent cancers: colorectal cancer, breast cancer, lung cancer, bladder cancer, neck cancer, uterine cancer, non-Hodgkin's lymphoma, skin cancer, myeloid leukemia, prostate cancer. In order to explain adverse effects, we notice that chemotherapy acts on bone marrow where white blood cells that regulate the body's response to infections, are produced. Neutrophils are specific white blood cells that often decrease after each chemotherapy session, which leads to febrile neutropenia. However, it is necessary to consider that a very low level of neutrophils could lead doctors to reduce certain doses of treatment and even to cancel an entire therapy cycle [5]. In addition, several prescription of growth stimulus can treat a neutropenia. Hemoglobin drop can train anemia with different symptoms such as fatigue, facial pallor, dizziness, respiratory difficulties, etc. Generally, it is necessary to regularize the treatment for anemia correction. Moreover, chemotherapy is responsible for feelings of nausea and vomiting which may occur on the same day or later in the treatment process [6].

In our work, we mainly consider side effects that are mostly stated after a chemotherapy session of about one week. Then, we focus on a multi-classification of the complication risk according to different severity levels. We notice that several national health institutes consider four severity levels for adverse events [7]: low, medium, severe, and highly severe. In our research and according to our study case, we only focus on three severity levels by preserving the two first ones (low, medium) and by combining the two other sublevels into only one level (severe). The last level necessitates fast medical interventions by adjusting drugs doses, reviewing their combination, prescribing new medications, etc.

The proposed support approach provides a web application for cancer patients in order to introduce information about side effects, one week after the last chemotherapy session. The provided information are analyzed through a machine learning module to predict the toxicity level of the patient. So, emergency and critical health states would be early identified in order to ensure fast medical intervention and to prevent any further complication.

In our approach, machine learning module is trained on stored medical data and acts on input data regarding patient health status. The outcome of this module is the severity level that indicates the complication risk of health for patient while undergoing chemotherapy treatment. According to the detected severity level, treating physician will decide which medical intervention is required in time. Consequently, the proposed approach will contributes to alert doctors to the patient’s conditions, which could be critical during the chemotherapy process.

In order to enhance the performances of the machine learning module, we propose different learning techniques that we adjusted and trained on medical data after a preprocessing step. During a test phase, we perform a comparative study of these techniques to decide which is the most efficient according to the most common evaluation metrics.

This paper is structured as follows. In section II, we present a review of machine learning applications for clinical prediction, especially in the area of cancer research. At the end of this review, we outline challenging issues that have not been addressed in literature such as risk complication and toxicity prediction in chemotherapy. In section III, we introduce the chemotherapy process and the resulting complication risks. The proposed support approach for toxicity prediction is detailed in Section IV. Theoretical aspects of the proposed machine learning module are introduced in section V. Evaluation metrics we consider for assessing the performances of the different machine learning techniques are discussed in section VI. The collected data about patients that are undergoing chemotherapy are explained in section VII. Then, in section VIII, we present the process of data preparation. Afterwards, we detail the prediction phase by considering the different learning models in section IX. An evaluation phase is finally, conducted according to performance metrics to decide which learning model performs better for our approach.

II. Literature Review

Given the importance of personalized medicine and the rising trend on applications of artificial intelligence techniques, we discuss in this section, the existing prediction models applied in healthcare systems and mainly in cancer research. In fact, to provide efficient and valuable healthcare services, an important component should be incorporated in healthcare systems to detect risks. With the massive volumes of regularly collected data, it is possible to identify future risks through predictive models by using machine learning techniques [8], [9].

In this context, several predictive models have been developed and used in healthcare systems. Nevertheless, for a better visibility of patient state and eventual risks, more advanced analytics are employed. Prediction models can be employed to determine patients at risks and to facilitate the management of healthcare process. They are also useful to carry out a risk adjustment by considering patient severity. Furthermore, these prediction models are valuable to identify successful or unsuccessful treatment.

The literature on prediction models in clinical and practical studies is very rich. The work presented in [10] provides a detailed review of early development and applications of medical prediction models. This review presents applications of statistical concepts, regression methods and strategies for developing and validating prediction models in public health, clinical practice, and medical research. In case of public health, prediction models are used in order to target preventive interventions for subjects at high risk, or for subjects developing a disease. Nevertheless, in clinical practice, the aim of prediction models is to inform patients and their treating doctors on the probability of diagnosis or prognostic outcomes. Moreover, authors in [11] proposed big data analytics for identifying patient with high risk and high healthcare cost. We also find the works of [12] who proposed a clinical prediction model in order to measure risks of hospitalizations for patients undergoing chemotherapy with advanced cancer. Medical records were used to abstract putative risk factors: patient characteristics, pretreatment values, treatment characteristics, etc. The proposed prediction model is based on multivariable logistic regression.

Authors in [13] proposed a transparent reporting of multivariable prediction models for individual prognosis or diagnosis known as TRIPOD. In [14], the authors detailed the construction of various clinical prediction models: logistic regression model, multiple linear regression model and Cox regression model. Efficiency of prediction models has to be assessed according to statistical analysis.

According to this brief review, the existing clinical prediction models rely on various regression models and data analytics. Further details about regression-modelling strategies are deeply discussed in [15]. The author provides technical details about linear models, logistic and ordinal regression, and survival analysis. The development of such models needs exhaustive expert efforts to collect data and to define significant features [16]. Recently, with the huge volumes of healthcare data and the increasing interest for healthcare system [17], the application of machine learning methods has become a significant emerging field. In fact, researchers and clinicians have applied different machine learning techniques for solving various problems of clinical outcomes prediction [18]. In [16], the author provides an overview of machine learning applications for clinical prediction tasks.

Machine learning approaches have been widely used in cancer research. First applications of these approaches refer to cancer predictive models [19]. Most of early applications involved different methods, such as Support Vector Machine, Decision

Trees, Neural Networks, and Bayesian Networks. These applications were dedicated for detecting, classifying tumors [4]. Afterwards, a particular focus on cancer prediction and prognosis, which are concerned with three prediction types: cancer susceptibility (or risk assessment), cancer recurrence and mortality [4], [20]. In [21] the authors provide a recent systematic review of learning based approaches for cancer prediction and diagnosis. In recent research [22], the authors used different learning models in order to assess the predictability of major cancer surgical outcomes while increasing the accuracy of previous traditional risk scores.

In [23], authors provide a review of the applied learning models to predict and classify Radiotherapy complications from two points of view: methodological and clinical. The authors focus on the type of the considered features as well as the used prediction methods with the main results. This overview involves published research about multiple cancer types (brain, head and neck, liver, breast, prostate esophagus and gynecological cancers).

Authors in [24], give a systematic review of machine learning methods that have been applied in cancer research for patient diagnosis, classification and prognosis. These methods involves reinforcement learning and deep learning techniques. A focus on deep learning based methods is also provided in a recent review [25] for breast cancer diagnosis. As a result of this review, Convolutional Neural Networks are the most accurate and extensively used method for breast cancer detection. As stated in [26], machine learning is increasingly applied in clinical oncology for various purposes varying from cancers diagnosis, outcomes predictions, to treatment design. In this context, the authors present recent advances of machine learning applications to clinical oncology. Moreover, a comparative study of widely applied machine learning methods in cancer detection is presented in [27]. This study only surveys research on the most common cancers worldwide: lung, breast, prostate and colorectal cancer.

According to the aforementioned reviews and research publications, machine learning methods had let to efficient and accurate decision making in healthcare field. In fact, these artificial intelligence based approaches are of a great help to improve the basic understanding of cancer development and progression [28].

However, in cancer research some particular concerns are still challenging. Among these issues, we mainly mention chemotherapy response and complications risks. In literature, few works have addressed the prediction of cancer therapy. Authors in [29], provide an overview of the latest progresses in therapeutic response prediction. They also discussed the use of machine learning algorithms and highlighted the recent challenges in therapy response prediction for clinical practice. In [2], the authors provides an overview of challenges and advances in drug response prediction. The authors also focus on comparing machine learning methods to be of greatest practical use for clinicians and artificial intelligence non-experts. In the works of [30], artificial intelligence methods and supervised learning techniques have been employed for drug repurposing in cancer. The objective in this work is to identify new therapeutic purposes for approved drugs after phenotypic observations.

Nevertheless, the discussed research works have not considered toxicity prediction and risk complication during chemotherapy. Therefore, in our work we focus on the prediction of toxicity level for cancer patient while undergoing chemotherapy. This toxicity prediction will be based on machine-learning methods as very promising research topic in medical field.

III. Chemotherapy Complications

In this section, we present one of the most common clinical practices that are applied in the medical sector for cancer treatment: chemotherapy. We also discuss its side effects and the necessity of treatment control to prevent high toxicity level. The main objective of chemotherapy is killing cancerous cells with the help of drugs, which quickly target dividing cells. Hence, the aim of these medications is to shrink tumors through different therapeutic levels. However, these medications have dangerous side effects [21], [31]:

- Immunotherapy: it boosts and helps the immune system to find cancer cells and attack them. Its side effects may be skin reactions and fatigue.
- Hormone-level therapy: Hormones play significant role among patients suffering from breast or prostate cancers. The side effects of hormonal therapy may be fatigue, nausea and vomiting, diarrhea, muscle pain.
- Molecularly targeted therapy: it uses drugs to target specific molecules. It aims to improve immunity by avoiding the spread and growth of cancer. Monoclonal antibodies and small-molecule drugs are some instances of target therapies. Side effects may include skin problems, high blood pressure and blood clots.
- Personalized medication: This newly developed approach determines suitable treatments for some specific cancer with the help of genetic tests.
- Radiation therapy: It works by destroying cancer cells and damaging a cancer cell’s DNA. So, it stops cells growing and dividing. Side effects depend on the type of radiation therapy, the dose of radiation and treatment schedule. Fatigue, loss of appetite, skin problems, anemia, loss of hair, nausea and vomiting are side effects of radiotherapy.

As explained in this section, side effects may happen with any kind of treatment. They mainly depend on the type of drug, the combination of drugs, the combination of treatment, the dose and the overall health. Some effects tend to be mild and disappear once the body gets used to the drug. However, when certain side effects are severe, treating physicians may decide to adjust the dose, adjust the combination of treatment, stop the therapy for a period of time or even change the drugs. Consequently, treating physicians need to be informed about the evolution of the therapy in order to provide the necessary medical help in time and avoid any further complications.

As we mentioned before, many National Health Institutes consider five severity levels for side effects [7]: low, medium, severe, highly severe and death related to adverse events. According to the collected data set, we only consider here three severity levels by preserving the two first ones and combining the remaining sublevels into only one level. Hence, Low level, Medium level and severe level are the considered toxicity levels. The last level corresponds to an emergency and requires urgent medical intervention in order to prevent worsening of health situation.

IV. Proposed Clinical Support Approach

In this work, our objective is to support medical decision processes for cancer patients who are undergoing a chemotherapy. As shown in Figure 1, we propose a clinical support approach that provides a web application for cancer patients in order to introduce information about side effects, one week after the last chemotherapy session. Thus, only severe and persistent side effects will be considered by treating physicians. Then, on the base of machine learning techniques, the severity level of patient is detected for a possible emergency. The involved prediction models acts on collected medical data regarding patient health status.

The proposed approach provides a support for medical staff by analyzing the collected data and predicting the severity level of health in order to decide a fast medical intervention. The detected severity level of a cancer patient may be low, medium or high. The proposed tool will contributes to prevent complication risks during the chemotherapy process of a cancer.

The proposed machine learning models are based on stored medical data that will be preprocessed and which dimensions are reduced in order to reveal models and create robust analysis. Precision and accuracy metrics are involved in the assessment of the proposed machine learning models to select the optimal one.

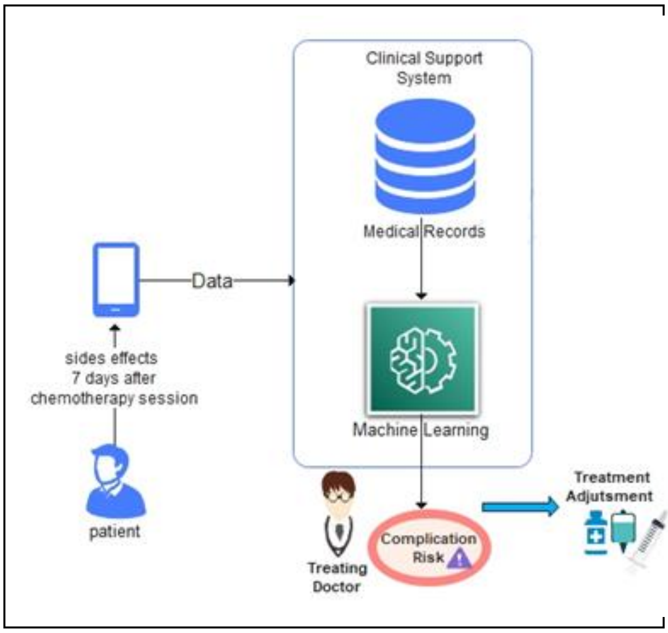


Figure 1. Clinical Support Approach for Toxicity Prediction

V. Toxicity Prediction Models

Since we focus on medical data analysis and prediction methods [8],[9], [32] we introduce in this section some useful theoretical concepts of the proposed prediction models. However, we firstly give a brief overview of machine learning process, which involves the following steps [9]:

- **Data Collection**: The key process in machine learning is collecting data from various sources.
- **Data Preparation**: Preprocessing data is a crucial step to purify the dataset and to obtain the optimized results. This preprocessing step involves data cleaning and features engineering. Data cleaning provides better understanding of the features and the relationships between them. Missing values and outliers in the collected dataset have to be handled to enhance the process and to improve the modeling performances. Essential variables have to be extracted and non-essential variables are removed. This task is big challenging and critical.
- **Model Building**: during this major step, learning model is trained to understand the outcomes. Then, a test of the model is performed by comparing the different outcomes. The training involves a percent of the data, while the remaining percent will be utilized to evaluate the model.
- **Model Evaluation**: This step aims to assess better fitment of model and data, and to compare different models for a correct model selection and accuracy in prediction.

Remember that we are interested here in supervised machine learning methods since we aim to predict a possible complication of patient's state according to category classification "Low", "Medium" and "High". Hence, our interest is mainly focused on Multi-class prediction. For each proposed model, we will explore its theoretical aspect and its hypotheses.

A. Linear Discriminant Analysis

LDA is one of the most common technique for dimension reduction, data visualization and classification (binary or multiclass). Its major advantages are simplicity, robustness and its interpretable classification results [33]. This linear transformation technique is commonly used for dimensionality reduction in the pre-processing phase for pattern-classification and machine learning applications. It makes predictions based on estimating the probability that a new set of inputs belongs to every class. Thus, the class with the highest probability is considered as the output class. Bayes theorem is implied in this model in order to estimate the probabilities.

LDA model estimates the mean and the variance of all input data for each class as follows [33]:

The mean μ of each input *x* for each class *k* can be normally estimated by dividing the sum of values by their total number:

$$m_k = \frac{1}{n_k}\sum x \qquad (1)$$

Where m_k is the mean value of *x* for class *k* and n_k the number of instances for class *k*.

The variance is calculated for all classes as the quadratic difference of each value *x* relative to the mean:

$$\sigma^2 = \frac{1}{n\ -k}\sum(x-\mu)^2 \qquad (2)$$

With σ^2 the variance of all input data x, n the number of instances, k the number of classes and μ the mean of each input x.

B. Naïve Bayes Model

Naive Bayes is a powerful algorithm for predictive modeling. This supervised learning technique is based on applying Bayes theorem and is usually employed for classification problems. This probabilistic machine learning model performs fast predictions to discriminate different objects based on their probabilities. Therefore, this model is one of the most efficient classification algorithms [34].

Bayes' rule or Bayes' theorem depends on conditional probability. In fact, this theorem is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability. Bayes' theorem is formally defined as [34]:

$$P(A|B) = (P(B|A) \times P(A)) / P(B) \qquad (3)$$

where:

- $P(A|B)$: Posterior probability means a probability of hypothesis A on the observed event B.
- $P(B|A)$: Likelihood probability, which corresponds to a probability of the evidence given that the probability of a hypothesis is true.
- $P(A)$: Prior Probability, which represents a probability of hypothesis before observing the evidence.
- $P(B)$: Marginal Probability, which corresponds to a probability of Evidence.

The main advantage of Naïve Bayes is its quick learning algorithm to predict a class of datasets. It can also be applied either for binary or multi-class classifications. Let us notice that in most applications of multi-class predictions, this model performs better when compared to other algorithms.

C. Decision Trees Model

Decision Trees are rule-based method for dealing with many classification and Regression problems. This model is a non-parametric method that does not rely on probability distribution assumptions. Its main advantage is its capacity to handle high dimensional data with good accuracy [32]. In this model, internal nodes correspond to features, branches represent decision rules, and leaf nodes are the outcomes. The root node learns to partition the tree based on the feature value. This partitioning is performed recursively (see Figure 2).

This classification technique states that the dataset features are tested from tree nodes. Then, for each possible value of the feature, a new branch is formed. This algorithm is performed until no variables have to be tested and leaf nodes are obtained to represent the different target classes [32]. Another advantage of this model is that it shares internal decision-making logic, contrarily to other algorithms like Neural Network (black-box type). Moreover, the required training time in this model is faster in comparison to neural network algorithm [32].

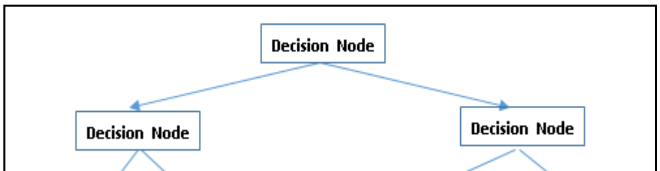


Figure 2. Decision Trees Model

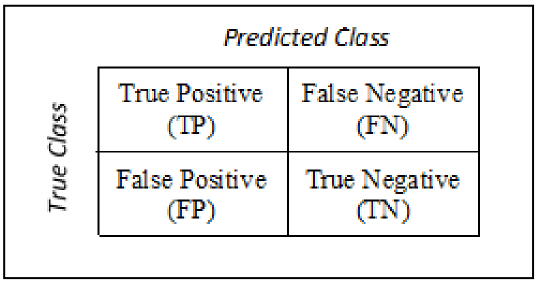
VI. Performance Metrics

In supervised machine learning, confusion matrix is a very popular and an important performance measurement to determine the quality of the classification system. It can be used to either binary or multi-class classification problems [35].

This matrix is used to assess the performance of classification models for a given validation dataset. Hence, the main requirement is to have a test dataset with expected outcome values. The matrix summarizes prediction results on a classification problem [35].

In fact, the number of correct and incorrect predictions are summarized through count values and broken down by each class. Thus, it indicates the ways in which the classification model is confused when predictions are made. Notice that the matrix is divided into two dimensions: predicted values and actual values along with the total number of predictions.

The predicted values are those generated by the model while actual values are the true ones for the given observations. Figure 3 shows the structure of confusion matrix for binary classification.



	Predicted Class	
True Class	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

Figure 3. Confusion Matrix for binary classification

This matrix shows the correctly classified TP values and FP values in the relevant class. It also represents the correctly classified TN values in the other class as well as FN values. On the base of this matrix, many performance metrics for classification may be calculated. The most frequently used metrics are accuracy, sensitivity, specificity, and Kappa Coefficient.

- **Accuracy** is the most important parameters to indicate the accuracy of the classification problems. It estimates how often the model gives true predictions. It is calculated as the ratio of the number of correct predictions made by the classifier to all number of generated predictions by the classifiers [35].

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \qquad (4)$$

- **Sensitivity**: it is defined as True Positive rate and corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points. Formally, it is calculated as follows :

$$Sensitivity = \frac{TP}{(TP + FN)} \qquad (5)$$

- **Specificity**: it is defined as True Negative rate and corresponds to the proportion of negative data points that are correctly considered as negative, with respect to all negative data points. Formally, it is calculated as follows:

$$Specificity = \frac{TN}{(FP + TN)} \qquad (6)$$

- **Kappa Coefficient**: it is an evaluation metric, which compares an observed accuracy with an expected accuracy (or random chance). The aim of this metric is to assess the performance of any classifier in relation to a “random classifier” [36]. Formally, it is defined as follows:

$$K = \frac{(Accuracy - randAccuracy)}{(1 - random)} \qquad (7)$$

The calculation of random accuracy *randAccuracy* is described as follows: from the confusion matrix, there is a randomly drawn label from the dataset that would be positive with probability p_1 and negative with probability $(1-p_1)$, with:

$$p_1 = \frac{(TP + FN)}{(TP + FP + TN + FN)} \qquad (8)$$

In addition, the considered classifier produces a positive label with probability p_2 and a negative label with probability $(1-p_2)$, with:

$$p_2 = \frac{(TP + FP)}{(TP + FP + TN + FN)} \qquad (9)$$

Random accuracy is the probability that the generated labels by these two processes accidently match:

$$randAccuracy = p_1 \times p_2 + (1 - p_1) \times (1 - p_2) \qquad (10)$$

Notice that when accuracy is 1, then Kappa coefficient K is equal to 1. Commonly, kappa coefficient captures a relative progress towards perfection from a random baseline. For instance, when random accuracy is estimated to 30% and the classifier accuracy is equal to 65% then, K is equal to 0.5. Hence, the current classifier is about 50% of the way to perfect from trivial performance.

VII. Cancer Patient Dataset

As we mentioned before, our case study is related to cancer patients that are undergoing chemotherapy. These patients have declared adverse effects after one week of a chemotherapy session. In the collected dataset, we considered ten types of frequent cancers: colorectal cancer, breast cancer, lung cancer, bladder cancer, neck cancer, uterine cancer, non-Hodgkin’s lymphoma, skin cancer, myeloid leukemia, prostate cancer. In table 1, we illustrate the dataset variables of this dataset [37].

Table 1. Description of Dataset variables

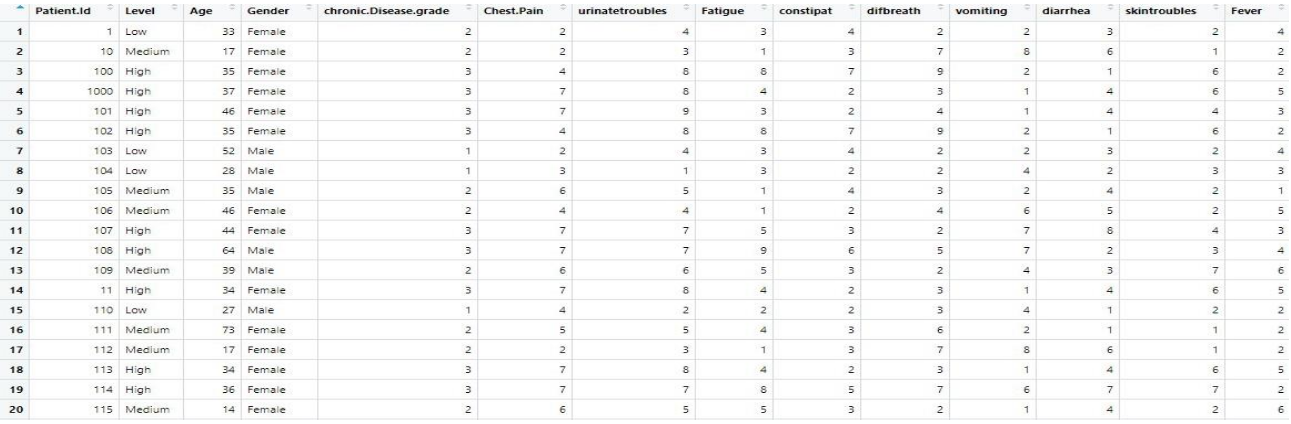
Dataset Variables	Description and values
Patient.ID	Unique patient identifier
ToxicityLevel	Target variable : High, Medium, Low
Age	Patient Age
Gender	Male or Female
Chronic Disease grade	Patient’s cancer stage : 1, 2, 3
Chest Pain	Level of Chest pain : in [1, 9]
Urinate troubles	Urinary tract infections: in [1, 9]
Fatigue	Tiredness signs: in [1, 9]
Constipation	Constipation Signs: in [1, 9]
DiffBreath	Level of breathing difficulties: in [1, 9]
Vomiting	Level of vomiting signs: in [1, 9]
Diarrhea	Level of Diarrhea signs: in [1, 9]
SkinTroubles	Level of skin troubles: in [1, 9]
Fever	Level of detected fever: in [1, 9]

As explained in Table 1, the dataset involves 13 input features and a target variable, which is *Toxicity Level.* Among input features, we consider those related to occurring symptoms: chest pain, urinate troubles, fatigue, constipation, difficulty in breathing, vomiting, diarrhea, skin troubles and fever.

The level of each detected symptom is defined within the interval [1,9] during the treatment process. In fact, through this level we indicate the intensity level of adverse events as well as an important parameter, which is the event occurrence.

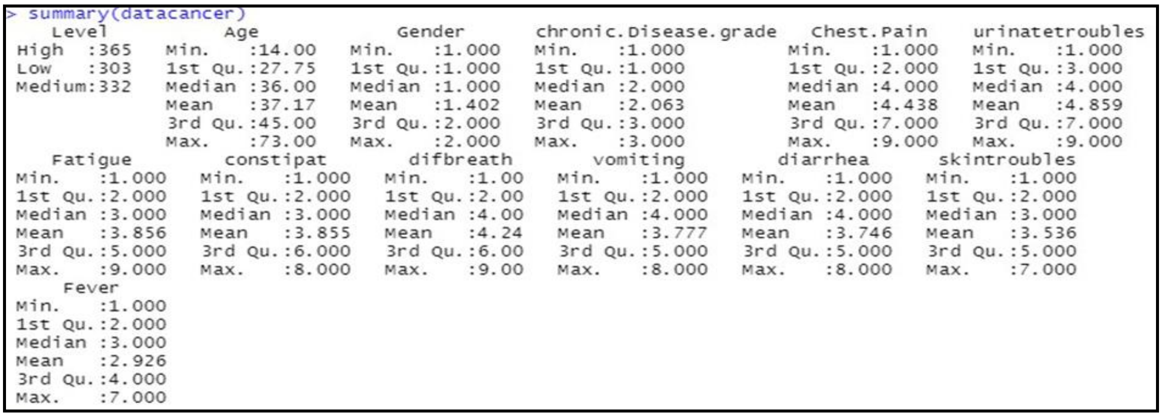
We consider for intensity level 1, the values 1, 2 and 3 which are assigned to a symptom when the side effect occurs once, twice or more times respectively, within a week.

For intensity level 2, the values 4, 5 and 6 are assigned when the side effect occurs once, twice or more within a week. Similarly, the values 7, 8 and 9 are considered according to the frequency of events of intensity level 3 within a week. In Figure 4, we illustrate a fragment of our dataset with the discussed variables. In order to summarize the available dataset, we opted for a descriptive statistic through some parameters. In Figure 5, we show the statistical parameters of each described variable for the considered population: minimal and maximal values, mean and average value. The most common issue in classification cases is the unbalanced distribution of classes. In fact, numerous classification algorithms requires accuracy for building predictive models in order to avoid the risk of dominant class in the result. In our case, the distribution of classes for the considered population is quite balanced as shown in Figure 6. Hence, the use of accuracy metrics will be significant for the evaluation of the proposed classification models.



	Patient.Id	Level	Age	Gender	chronic.Disease.grade	Chest.Pain	urinatetroubles	Fatigue	constipat	difbreath	vomiting	diarrhea	skintroubles	Fever
1	1	Low	33	Female	2	2	4	3	4	2	2	3	2	4
2	10	Medium	17	Female	2	2	3	1	3	7	8	6	1	2
3	100	High	35	Female	3	4	8	8	7	9	2	1	6	2
4	1000	High	37	Female	3	7	8	4	2	3	1	4	6	5
5	101	High	46	Female	3	7	9	3	2	4	1	4	4	3
6	102	High	35	Female	3	4	8	8	7	9	2	1	6	2
7	103	Low	52	Male	1	2	4	3	4	2	2	3	2	4
8	104	Low	28	Male	1	3	1	3	2	2	4	2	3	3
9	105	Medium	35	Male	2	6	5	1	4	3	2	4	2	1
10	106	Medium	46	Female	2	4	4	1	2	4	6	5	2	5
11	107	High	44	Female	3	7	7	5	3	2	7	8	4	3
12	108	High	64	Male	3	7	7	9	6	5	7	2	3	4
13	109	Medium	39	Male	2	6	6	5	3	2	4	3	7	6
14	11	High	34	Female	3	7	8	4	2	3	1	4	6	5
15	110	Low	27	Male	1	4	2	2	2	3	4	1	2	2
16	111	Medium	73	Female	2	5	5	4	3	6	2	1	1	2
17	112	Medium	17	Female	2	2	3	1	3	7	8	6	1	2
18	113	High	34	Female	3	7	8	4	2	3	1	4	6	5
19	114	High	36	Female	3	7	7	8	5	7	6	7	7	2
20	115	Medium	14	Female	2	6	5	5	3	2	1	4	2	6

Figure 4. Fragment of Cancer Patient Dataset



```
> summary(datacancer)
    Level          Age            Gender       chronic.Disease.grade   Chest.Pain     urinatetroubles
 High  :365   Min.   :14.00   Min.   :1.000   Min.   :1.000          Min.   :1.000   Min.   :1.000
 Low   :303   1st Qu.:27.75   1st Qu.:1.000   1st Qu.:1.000          1st Qu.:2.000   1st Qu.:3.000
 Medium:332   Median :36.00   Median :1.000   Median :2.000          Median :4.000   Median :4.000
              Mean   :37.17   Mean   :1.402   Mean   :2.063          Mean   :4.438   Mean   :4.859
              3rd Qu.:45.00   3rd Qu.:2.000   3rd Qu.:3.000          3rd Qu.:7.000   3rd Qu.:7.000
              Max.   :73.00   Max.   :2.000   Max.   :3.000          Max.   :9.000   Max.   :9.000
    Fatigue        constipat       difbreath       vomiting        diarrhea       skintroubles
 Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
 Median :3.000   Median :3.000   Median :4.00   Median :4.000   Median :4.000   Median :3.000
 Mean   :3.856   Mean   :3.855   Mean   :4.24   Mean   :3.777   Mean   :3.746   Mean   :3.536
 3rd Qu.:5.000   3rd Qu.:6.000   3rd Qu.:6.00   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000
 Max.   :9.000   Max.   :8.000   Max.   :9.00   Max.   :8.000   Max.   :8.000   Max.   :7.000
     Fever
 Min.   :1.000
 1st Qu.:2.000
 Median :3.000
 Mean   :2.926
 3rd Qu.:4.000
 Max.   :7.000
```

Figure 5. Statistical Parameters of variables in the dataset

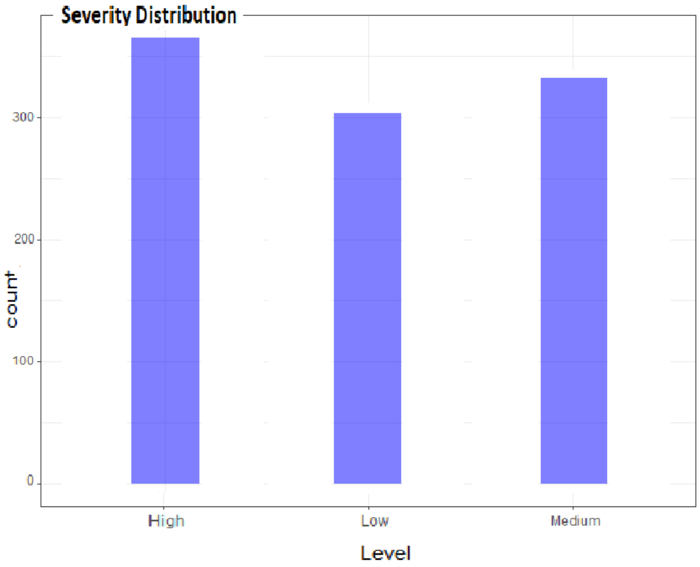


Figure 6. Distribution of classes in the dataset

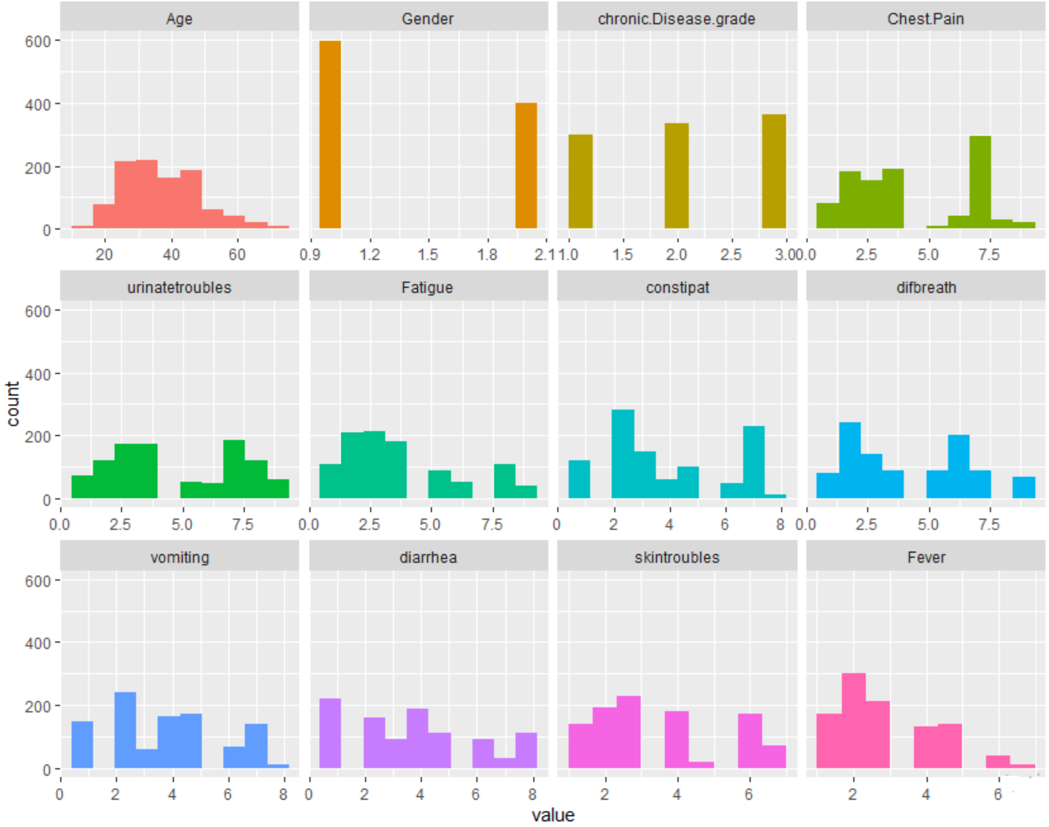
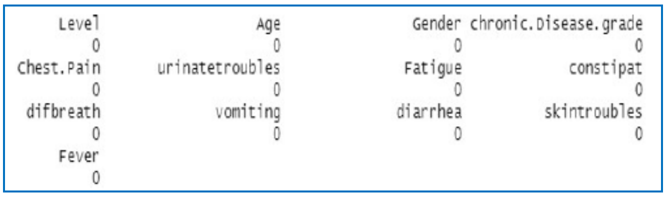


Figure 7. Distribution of important variables in the dataset

Figure 7 illustrates the distribution of the most important variables in the dataset, without considering PatientID and the target variable Toxicitylevel. We notice that the variable Age follows normal distribution. For the remaining variables, we observe single bars of histograms, which correspond to outliers. Therefore, in order to decrease risks of errors, a preprocessing step is firstly required to prepare the training dataset through a process of cleaning and correcting the raw dataset. Let us notice that in our approach, the overall dataset is organized into two parts: Training dataset (80% of the overall dataset) and Test dataset (20% of the overall dataset). Therefore, we opted for the cross-validation method, which uses a resampling procedure to split the dataset into two parts. This method is employed on different iterations in order to evaluate the proposed machine learning models by using different portions of the data to train and test each model.

VIII. Data Preparation

As a first step of data preprocessing, an inspection of the raw dataset for missing values is required. The result of this inspection is illustrated in Figure 8. Since machine learning models suppose the independence of predictive variables, it is necessary to check correlations between the different variables. According to the generated correlation matrix, some correlations between the different variables are observed [37]. So, it is required to remove the identified correlations. For that purpose, we opted for the Principal Component Analysis (PCA) which is the most common technique in literature for transforming a set of features into smaller one [38].



```
      Level                  Age               Gender chronic.Disease.grade
          0                    0                    0                     0
 Chest.Pain      urinatetroubles              Fatigue             constipat
          0                    0                    0                     0
  difbreath             vomiting             diarrhea          skintroubles
          0                    0                    0                     0
      Fever
          0
```

Figure 8. Inspection for missing values

PCA is a linear dimensionality reduction technique. Its main purpose is the extraction of information from high-dimensional set by projecting it into a lower-dimensional sub-set. This transformation should preserve the important parts with more variation of the data while removing the non-essential parts that have fewer variations of the data. In our case, we applied the PCA in order to transform the correlated variables into new uncorrelated Principal Components (PC).

We notice that the reason of generating uncorrelated PC from the original variables is that the correlated variables contribute to the same principal component. Thereby, this feature reduction lead to uncorrelated principal components, where each one corresponds to a different set of correlated features that have different amounts of variation [39].

Nevertheless, a challenging issue of dimensionality reduction to take into account is the trade off between accuracy and simplicity.

In order to determine the number of Principal Components *p*, from a set of features, there are different approaches. First approach consists in retaining components with variance more than 0.1, since these components have interpretive value.

Another approach to determine *p*, is the visual inspection of a scree plot diagram. This visualization shows the explained variance (or eigenvalues) in a downward curve according to decreasing order of eigenvalues. The elbow of the graph where the explained variances seem to stabilize is determined. Then, we preserve the components in the left side of this point as important factors [30]. This approach is explained by statistical data through multivariate normal distribution with correlations of 0.0. In Table 2, we show the statistical data about the importance of components before dimensionality reduction: standard deviation, proportion of variance and cumulative variance.

Table 2. Importance of all components before applying PCA

	Standard Deviation	**Proportion of Variance**	**Cumulative Proportion**
PC1	1.8972	0.3272	0.3272
PC2	1.2915	0.1516	0.4789
PC3	1.2337	0.1384	0.6172
PC4	1.0694	0.1040	0.7212
PC5	0.9755	0.0865	0.8077
PC6	0.81958	0.06106	0.86877
PC7	0.74520	0.05048	0.91925
PC8	0.57555	0.03011	0.94937
PC9	0.49467	0.02225	0.97161
PC10	0.44722	0.01818	0.98980
PC11	0.3350	0.0102	1.0000

According to this analysis, the first two components explain ≈ 0,48 of the variance (information quantity). We also notice that eight principal components (PC1 → PC8) are required to explain more than ≈ 0.94 of the information quantity. Moreover, ten components are needed to explain more than ≈ 0.98. Figure 9 shows the scree plot before data reduction. We notice that for each component, the variance proportion is more than 0.1.

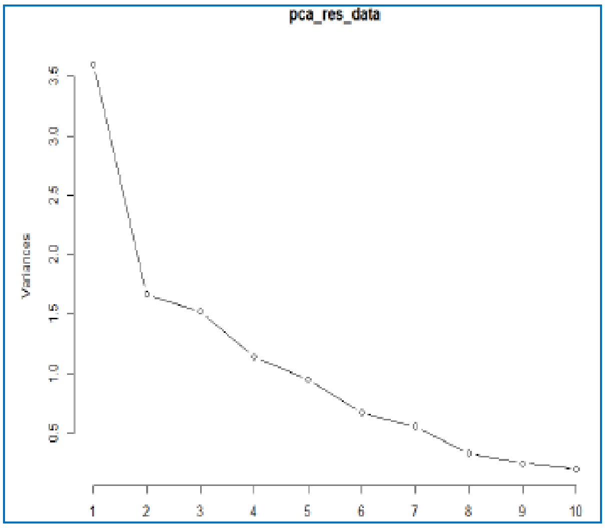


Figure 9. Scree Plot before before reduction

The scree plot starts to stabilize after the 8th component so that we should focus on the eight first components. Hence, after dimensionality reduction process by using PCA, we obtain the results in Table 3, which illustrates the importance of principal components after data reduction. The obtained results illustrate that 7 principal components explain 97% of the variance in the transformed dataset. The generated scree plot of Figure 10, confirm these results.

Table 3. Importance of all components after applying PCA

	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	1.495	0.2794	0.2794
PC2	1.2666	0.2005	0.4799
PC3	1.0435	0.1885	0.6685
PC4	1.0435	0.1361	0.8046
PC5	0.80726	0.08146	0.88608
PC6	0.67711	0.05731	0.94339
PC7	0.50526	0.03191	0.9753
PC8	0.4445	0.0247	1.0000

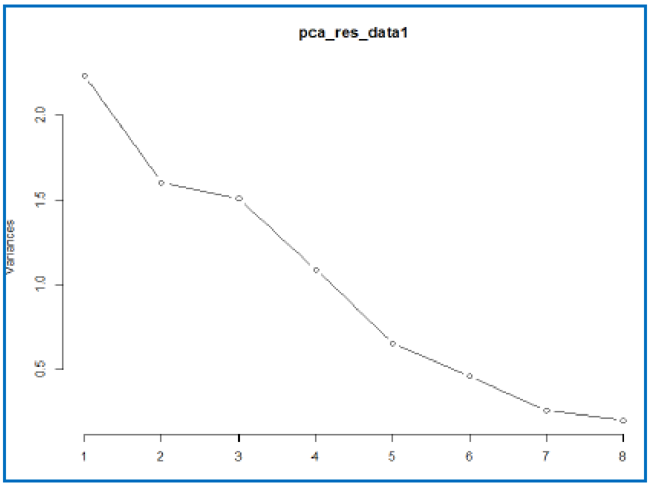


Figure 10. Scree plot after reduction

IX. Toxicity Prediction Module

After preparing the dataset from data cleaning to data reduction, the prediction process is then, performed. In this section, we present this process by using the discussed machine learning models: Naïve Bayes, Linear Discriminant Analysis, Decision tree.

Let us remember that in our case study, the overall dataset (1000 instances) is partitioned into two subsets:

- Training dataset which corresponds to 80% of the overall dataset
- Test dataset which represents 20% of the overall dataset

This dataset partitioning was performed by using the cross-validation method, which involves a resampling procedure to split the dataset into two parts. Cross-validation method is used on different iterations in order to evaluate the proposed machine learning models by using different portions of the data to train and test each model. In the remaining of this section, we expose the obtained results on the test parts (200 instances).

A. Naïve Bayes Model

After training our Naïve Bayes model on training data, we performed the test phase on the remaining data. According to the obtained results, the distribution of correct prediction is of 35% of the total number of cases. In Table 4, we illustrate the frequency of each true class as well as the frequency of each predicted class in the test dataset. The detailed statistical results of the prediction model according to each class are illustrated in Table 5.

Table 4. Obtained frequencies with Naïve Bayes Model

Class	Frequency of True Class	Frequency of Predicted Class
High	35.17%	36.68%
Medium	35.67%	30.15%
Low	29.14%	33.16%

Table 5. Statistical results for each class with Naïve Bayes Model

	High	Medium	Low
Sensitivity	0.9589	0.8788	1.0000
Specificity	1.0000	1.0000	0.9209
Positive Predictive Value	1.0000	1.0000	0.8451
Negative Predictive Value	0.9767	0.9433	1.0000
Prevalence	0.3668	0.3317	0.3015
Detection Rate	0.3518	0.2915	0.3015
Detection Prevalence	0.3518	0.2915	0.3568
Balanced Accuracy	0.9795	0.9394	0.9604

Sensitivity and specificity are illustrated for each class in order to indicate the ability of the model to correctly classify the patient.

According to these results, we notice that sensitivity and specificity are relatively high for the different classes. Thereby, few false negative results are found in class High and Low (sensitivity ≈ 96% and 100%). In the same way, few false positive results are found in the three classes (specificity ≈ 100% (High, Medium), 92% (Low)). Moreover, the accuracy of each class is relatively high since it is between 94% and 98%.

B. Linear Discriminant Analysis Model

According to the obtained results by using linear discriminant analysis, we notice that the model leads to correct predictions for 35% of the total number of cases. It also generates incorrect predictions for the other two classes with lower than 10%.

In Table 6, we summarize the frequency of each true class, as well as the frequency of each predicted class in the test dataset. Moreover, we illustrate in Table 7, the statistical results for each class with LDA model.

When comparing these results with those of Naïve Bayes Model, we find that sensitivity and specificity are lower. Thereby, false negative and more false positives results are more detected here. In addition, the accuracy for each class is less than the previous model since it is between 89% and 94%.

Table 6. Obtained Proportions with Linear Discriminant Analysis

Class	Frequency of True Class	Frequency of Predicted Class
High	39.69%	36.68%
Medium	29.14%	30.15%
Low	31.15%	33.16%

Table 7. Statistical results for each class with linear discriminant analysis Model

	High	Medium	Low
Sensitivity	0.9589	0.8333	0.9000
Specificity	0.9286	0.9474	0.9712
Negative Predictive Value	0.9750	0.9197	0.9574
Prevalence	0.3668	0.3317	0.3015
Detection Rate	0.3518	0.2764	0.2714
Detection Prevalence	0.3970	0.3116	0.2915
Balanced Accuracy	0.9437	0.8904	0.9356

C. Decision Trees Model

With Decision Trees model, we obtained correct predictions for 35% of the total number of cases. The frequency of each true class as well as for each predicted class in the Test dataset are summarized in Table 8. We also detail in Table 9, the obtained statistical results for each class with Decision Trees Model.

Table 8. Obtained Proportions with Decision Trees Model

Class	Frequency of True Class	Frequency of Predicted Class
High	39.69%	36.68%
Medium	38.69%	30.15%
Low	29.14%	33.16%

According to the obtained results, the tradeoffs between sensitivity and specificity are comparable to those resulting from Linear Discriminant Analysis. However, sensitivity and specificity resulting from Naïve Bayes are clearly higher for each class.

Table 9. Statistical results for each class with Decision Trees Model

	High	Medium	Low
Sensitivity	0.9589	0.6515	0.9667
Specificity	0.9286	1.0000	0.8633
Positive Predictive Value	0.8861	1.0000	0.7532
Negative Predictive Value	0.9750	0.8526	0.9836
Prevalence	0.3668	0.3317	0.3015
Detection Rate	0.3518	0.2161	0.2915
Detection Prevalence	0.3970	0.2161	0.3869
Balanced Accuracy	0.9437	0.8258	0.9150

Let us also notice that the obtained accuracy for each class is lower than the results of LDA, since it is between 83% and 94%. These results are clearly lower than those of Naïve Bayes Model.

D. Selection of Learning Model

After training and testing the machine learning models for our case study, an assessment of performances is required. This comparative study is carried out on the base of performance metrics: accuracy and Kappa coefficient. In fact, since the distribution of the dataset according to the different classes is balanced, the accuracy metric will be the most significant one. We also consider the Kappa coefficient, which assess the performance of any classifier in relation to a random one as explained in this paper. In Figure 11, we illustrate the performances of the three models according to accuracy. As we can notice, Naïve Bayes model achieved the best performance with accuracy of ≈ 95%, in comparison to Linear Discriminant Model with accuracy ≈ 90% and Decision Tree model with accuracy ≈ 86%.

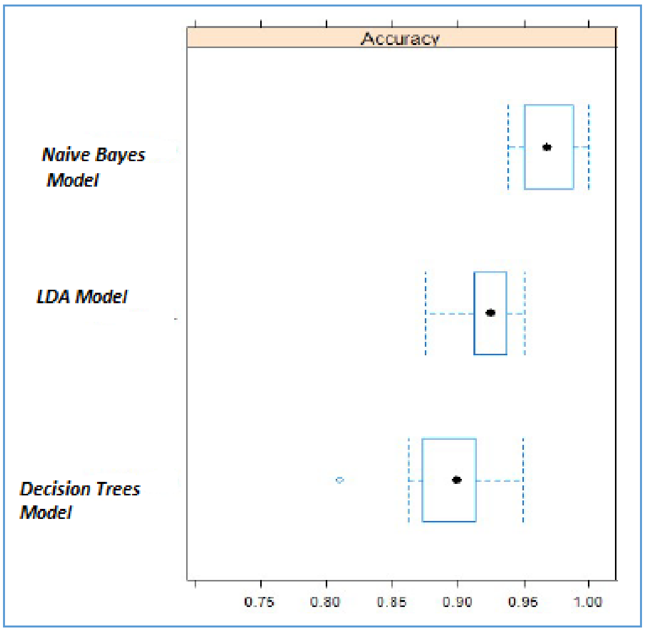


Figure 11. Performances of the ML Models according to Accuracy metric

When considering the Kappa coefficient, we obtained the comparable results as illustrated in Figure 12.

This coefficient estimates the quality of predictions. In fact, it describes the proportional reduction of prediction error for a classification model. Then, it compares it to the generated error of a completely hazardous classification. In other words, when Kappa coefficient is equal to 0.9, this result means that 90% of the obtained classification are not hazardous results. In our case when applying this coefficient, we obtained the following values: ≈ 79% for Decision Trees Model, ≈ 92% for Naïve Bayes Model and ≈ 85% for LDA Model.

Details about the obtained values of performances metrics for the different machine learning models are summarized in Table 10.

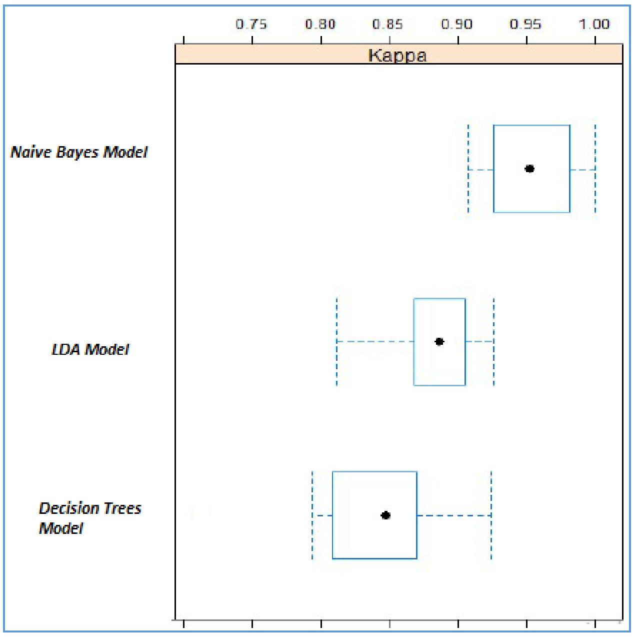


Figure 12. Performances of the ML Models according to Kappa Coefficient

Consequently, in our case study the best performances of predictions have been reached by using Naïve Bayes model. Thereby, the prediction module in our support approach will be based on this model for an effective prediction of toxicity level [40].

Table 10. Summary of Performance Metrics for ML Models

	Naïve Bayes Model	LDA Model	Decision Trees Model
Accuracy	0.9447	0.8995	0.8593
95%CI	(0.9032, 0.9721)	(0.8491, 0.9375)	(0.8031, 0.9044)
No Information Rate	0.3668	0.3668	0.3668
P-Value [Acc > NIR]	< 2.2e-16	<2e-16	2.2e-16
Kappa	0.9171	0.8485	0.7887

X. Conclusions

Chemotherapy is the most common and effective treatment of many types of cancer. Nevertheless, like other cancer treatments, it frequently causes various side effects that may lead to complication risks for patient status. Thereby, patients should be aware of chemotherapy side effects so they could share them with the health care team.

In this scope, we proposed to support the medical staff during the decision process of chemotherapy through a monitoring and control system. Our approach is based on medical records that were collected from previous experiences of cancer patients during chemotherapy. In this data, we focus on the most common symptoms detected after the last chemotherapy session. On the base of this data, a machine learning module performs an early prediction of toxicity level, which may be low, medium or high. A high toxicity level indicates an emergency and a high complication risk for patient. For this purpose, we started by studying, refining and transforming the collected data. Then, we conducted the prediction process by using three prediction models: Naïve Bayes, Linear Discriminant Analysis and Decision Trees. The obtained results have shown the accuracy of the proposed learning models. The best classifications have been achieved by using Naïve Bayes, which outperforms the other two models with ≈95% of accuracy. Sensitivity and specificity metrics have also confirmed these results since they are higher in the case of Naïve Bayes. When considering the Kappa coefficient, which estimates the quality of predictions, we have also found that the best value was found with Naïve Bayes Model ≈ 92%.

References

[1] H. Lazaar-Ben Gobrane, S. Hajjem, H. Aounallah-Skhiri, N. Achour, M. Hsairi. “Mortalité par cancer en Tunisie : calcul des années de vies perdues”, *Revue de Santé Publique*, 1(23) pp. 31-40, 2011.

[2] G. Adam, L. Rampasek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, A. Goldenberg. “Machine learning approaches to drug response prediction: challenges and recent progress”. *NPJ Precis. Oncol*. 4(19), pp. 1-10, 2020. doi: 10.1038/s41698-020-0122-1.

[3] Statistiques Nationales sur les causes de décès en Tunisie. Ministère de santé, Institut National de santé, http://www.santetunisie.rns.tn, Rapport Avril 2021.

[4] K. Kourou, T.P., Exarchos, K.P. Exarchos, MV. Karamouzis, DI. Fotiadis. “Machine learning applications in cancer prognosis and prediction”. *Computational and Structural Biotechnology Journal* (13), pp.8–17, 2014. doi:10.1016/j.csbj.2014.11.005

[5] D. Cameron. “Management of chemotherapy-associated febrile neutropenia”, *British Journal of Cancer* 101(1), p.18–22, 2015. doi: 10.1038/sj.bjc.6605272

[6] C. Carr., J. Ng., T. Wigmore. “The side effects of chemotherapeutic agents”. *Journal of Current Anaesthesia & Critical Care* (19), pp.70-79, 2008. doi:10.1016/j.cacc.2008.01.004

[7] Common Terminology Criteria for Adverse Events (CTCAE). *Report Version 5.0.* November 27, 2017.

[8] F. Hutter, L. Kotthoff, J. Vanschoren. *Automated Machine Learning-Methods, Systems, Challenges*. Cham Springer Nature, 2019. doi : 10.1007/978-3-030-05318-5

[9] X.D. Zhang. “Machine learning”, in *A Matrix Algebra Approach to Artificial Intelligence*. Springer Singapore, pp. 223-440, 2020.

[10] EW. Steyerberg. *Clinical prediction models: A practical Approach to Development, Validation and Updating*. Springer, 2009. doi:10.1007/978-0-387-77244-8

[11] DW. Bates, S. Saria, L. Ohno-Machado, et al. “Big data in health care: using analytics to identify and manage high-risk and high-cost patients”. *Health Aff.* (Millwood) 33(7), pp. 1123-1131, 2014. doi:10.1377/hlthaff.2014.0041

[12] GA. Brooks, AJ. Kansagra, SR. Rao, et al. “A Clinical Prediction Model to Assess Risk for Chemotherapy-Related Hospitalization in Patients Initiating Palliative Chemotherapy”. *JAMA Oncology*, 1(4),pp.441-447,2015. doi:10.1001/jamaoncol.2015.0828

[13] GS. Collins, JB. Reitsma, DG. Altman et al. “Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement”. *Journal of European Urology*, 67(6), pp.1142-1151, 2014. doi:10.1016/j.eururo.2014.11.025

[14] ZR. Zhou, WW. Wang, Y.Li, et al. “In-depth mining of clinical data: the construction of clinical prediction model with R”. *Annals of Translational Medicine* 7(23), pp.796, 2019. doi:10.21037/atm.2019.08.63

[15] E. Frank, Jr. Harrell. “Ordinal logistic regression”, in *Regression modeling strategies*. Cham: Springer Series in Statistics, pp. 311-325, 2015.

[16] WH. Weng. “Machine Learning for Clinical Predictive Analytics”, in *Leveraging Data Science for Global Health*, L. Celi, M. Majumder, P. Ordóñez, J. Osorio, K. Paik K. and M. Somai (eds) Springer, Cham, 2020. doi: 10.1007/978-3-030-47994-7_12

[17] J. Henry, Y. Pylypchuk, T. Searcy, V. Patel. “Adoption of electronic health record systems among US non-federal acute care hospitals”, *ONC Data Brief* 35, pp. 2008-2015, 2016.

[18] M. Ghassemi, T. Naumann, P. Schulam, et al. “Opportunities in machine learning for healthcare”, 2018. arXiv:1806.00388.

[19] JA. Cruz, DS. Wishart. “Applications of machine learning in cancer prediction and prognosis”. *Cancer Informatics* 2, pp.59–77, 2007.

[20] D. Ding, T. Lang, D. Zou et al. “Machine learning-based prediction of survival prognosis in cervical cancer”. *BMC Bioinformatics* 22(331), 2021. doi:10.1186/s12859-021-04261-x

[21] Y. Kumar, S. Gupta, R. Singla, et al. “A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis”. *Arch. Computat. Methods. Eng.* 2021. Doi:10.1007/s11831-021-09648-w

[22] DM. Goncalves, R. Henriques L. Santos, R.S. Costa. “On the predictability of postoperative complications for cancer patients: a Portuguese cohort study”. *BMC Medical Informatics and Decision Making* 21(200), Springer, 2021. doi: 10.1186/s12911-021-01562-2

[23] L.J. Isaksson, M. Pepa, M. Zaffaroni et al. “Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy”. *Frontiers in Oncology* 10(790), 2020. doi: 10.3389/fonc.2020.00790.

[24] K. Kourou, KP. Exarchos, C. Papaloukas, P. Sakaloglou, T.Exarchos, DI. Fotiadis. Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. *Computational and Structural Biotechnology Journal*, 6(19), pp. 5546-5555, 2021. doi: 10.1016/j.csbj.2021.10.006.

[25] M. Nasser, U.K. Yusof. Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. *Journal of Diagnostics* 13(1), pp. 161, 2023. doi: 10.3390/diagnostics13010161

[26] K. Swanson, E. Wu, A. Zhang, A.A. Alizadeh, J. Zou. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment, *Journal of Cell*, 186(8), pp. 1772-1791, 2023. Doi: 10.1016/j.cell.2023.01.035.

[27] M. Mokoatle, V. Marivate, D. Mapiye, *et al.* A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application. Journal of *BMC Bioinformatics*, 24(112), 2023. doi: 10.1186/s12859-023-05235-x

[28] Y. Kumar, R. Singla. “Federated learning systems for healthcare: perspective and recent progress”, in *Federated learning systems*, MH. Rehman and MM Gaber (eds), Studies in computational intelligence 965, Cham: Springer, 2021. doi:10.1007/978-3-030-70604-3_6

[29] R. Rafique, S.M. Riazul Islam, J.U. Kazi, “Machine learning in the prediction of cancer therapy”. *Computational and Structural Biotechnology Journal* 19, pp.4003-4017, 2021. doi:10.1016/j.csbj.2021.07.003.

[30] Z. Tanoli, M. Vaha-Koskela, T. Aittokallio. “Artificial intelligence, machine learning, and drug repurposing in cancer”. *Expert Opin. Drug Discov*, pp. 1-13, 2021.

[31] Canadian Cancer Society, www.cancer.ca

[32] D. Che, Q. Liu, K. Rasheed et al. “Decision tree and ensemble learning algorithms with their applications in bioinformatics”, in *Software Tools and Algorithms for Biological Systems, Advances in Experimental Medicine and Biology* 696, pp. 191-199, 2011. doi:10.1007/978-1-4419-7046-6_19

[33] P. Xanthopoulos, PM. Pardalos, TB. Trafalis. “Linear discriminant analysis“, in *Robust data mining*. Springer, pp. 27-33, 2013.

[34] GI. Webb. “Naïve Bayes”, in *Encyclopedia of Machine Learning*. Springer, Boston, pp.713-714, 2011. doi:10.1007/978-0-387-30164-8_576.

[35] G. Shobha, S. Rangaswamy. “Machine learning”, in *Handbook of statistics* 38, Elsevier, pp. 197-228, 2018. doi:10.1016/bs.host.2018.07.004

[36] S. Uddin, A. Khan, M. Hossain et al. “Comparing different supervised machine learning algorithms for disease prediction”. *BMC Medical Informatics and Decision Making*, 19(281), Springer, 2019. doi:10.1186/s12911-019-1004-8

[37] I. Boudali, I. Belhadj Messaoud. “Machine Learning Models for Toxicity Prediction in Chemotherapy”, in *Intelligent Systems Design and Applications*, A. Abraham et al. (Eds.), *22nd International Conference on Intelligent Systems Design and Applications (ISDA 2022)*, December 12-14, 2022. Lecture Notes in Networks and Systems 646, Springer 2023. doi:10.1007/978-3-031-27440-4_202

[38] I. Jolliffe. “Principal Component Analysis”, in *International Encyclopedia of Statistical Science*, M. Lovric (eds), Springer, Berlin, Heidelberg, 2011. doi:10.1007/978-3-642-04898-2_455

[39] R. Bro, A.K. Smilde. “Principal Component Analysis”, in *Analytical methods*, 6(9), pp. 2812-2831, 2014. doi:10.1039/C3AY41907J

[40] L Dora, S Agrawal, R Panda, A Abraham, Optimal breast cancer classification using Gauss–Newton representation based algorithm, Expert Systems with Applications, 97: 134-145, 2017.

Author Biographies

Imen Boudali is currently an Assistant Professor in Computer Science at the University of Tunis El Manar, Tunisia. She received the Ph.D. degree in Computer Science from the University of Manouba in Tunisia in December 2009. She has almost twenty years of experience in the fields of higher education and scientific research. She is member of SERCOM Laboratory at the University of Carthage. Her research interests include intelligent systems, decision aid, optimization and machine learning.