International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Application of machine learning for analyzing cancer patient data and predicting survival

Diptarshi Mitra *

Global Institute of Health Science, Ahmedabad, India.

International Journal of Science and Research Archive, 2025, 14(01), 949-953

Publication history: Received on 24 November 2024; revised on 29 December 2024; accepted on 31 December 2024

Article DOI: https://doi.org/10.30574/ijsra.2025.14.1.2660

Abstract

Cancer is a deadly disease, and a leading cause of death globally. Thus, the prediction of the possibility of survival of cancer patients, at an early stage of treatment, will be beneficial for both the doctors and the patients. This study has attempted to predict the survival status of cancer patients, by employing two well-known Machine Learning algorithms viz., Logistic Regression and Support Vector Machine, and utilizing a dataset of Kaggle. Before using the Machine Learning models, suitable encoding and scaling techniques have been applied on the data. However, neither of the Machine Learning algorithms has performed satisfactorily (accuracy of prediction for Logistic Regression: 51.6%, and that for Support Vector Machine: 52.2%), and the actual reason for this poor performance seems to be the low quality and/or the insufficiency of the data used.

Keywords: Cancer Patient Survival; Logistic Regression; Support Vector Machine; Kaggle

1. Introduction

Data analysis is an important tool which helps in understanding the nature of a dataset, and extracting meaningful information from the dataset. Today's world is dominated by computer, Internet, and Machine Learning; almost every sphere of life is influenced by them. Likewise, the quality and the speed of data analysis can be enhanced by involving computer and Machine Learning. And, electronic data is easily available nowadays; Internet is also a source of data.

Medical data analysis is needed for finding out important information regarding the relevant patient/s and/or disease/s. And, computer, Internet, and Machine Learning can be utilized for analyzing medical data which may include cancer patient data. It is well-known that cancer is a deadly disease, and a lot of people die of cancer every year. So, both the doctors and the patients will be benefitted if the possibility of survival can be predicted at an early stage of treatment, by employing suitable Machine Learning models, and utilizing relevant data related to the patients, the tumors they have, and the treatments they have undergone so far. Then, alternative and better techniques of treatment may be planned for the patients having low probability of survival. And, the case histories of the patients with high probability of survival, may be investigated to gain new insights into the treatment of cancer.

Thus, the objective of this study is to predict the survival status of cancer patients by applying suitable Machine Learning models on electronic cancer patient data, obtained from the Internet.

1.1. Cancer

Cancer is a disease characterized by the uncontrollable growth and multiplication of some of the cells of the body. These abnormally growing and multiplying cells may form tumors (lumps of tissues). However, tumors may be cancerous or non-cancerous (benign). Cancerous tumors can spread into nearby and even distant regions of the body. It may be noted

* Corresponding author: Diptarshi Mitra.

further that cancer is a genetic disease i.e., it is caused by the changes in the genes which control the way of functioning of the cells, particularly, the way in which they grow and divide.

Many types of cancers are there. Some of the most common cancers are breast cancer, lung cancer, colon and rectum cancer, prostate cancer. Cancer is one of the leading causes of death worldwide. The image of a cancer cell is shown in fig.-1.

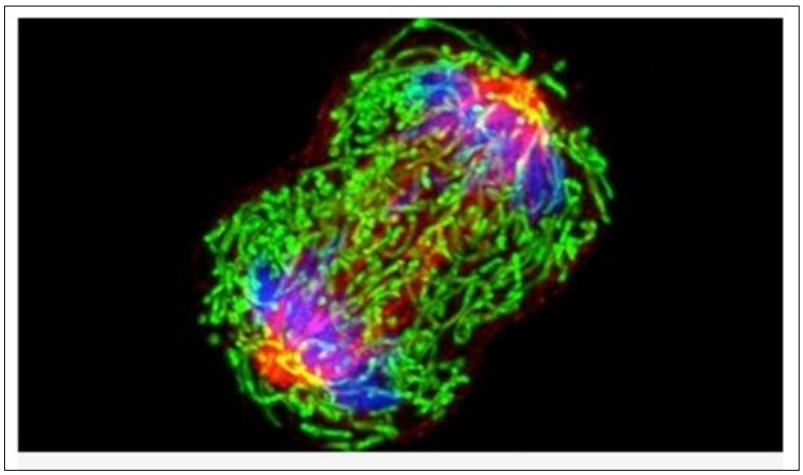

Figure 1 A dividing breast cancer cell [1]

1.2. Machine Learning

Machine Learning is the technique of programming computers to enable them to learn from data. In other words, it is the discipline which enables computers to learn without being explicitly programmed. Machine Learning is nowadays used in a large number of fields. It is particularly suitable for problems:

- where the solutions using the traditional approaches, involve a lot of hand-tuning and a large number of rules,
- where the traditional approaches cannot offer any suitable solution,
- which involve fluctuating environments, and
- where one needs a deep understanding of the nature of the problem and the data involved.

1.3. Data

The data used in this project, has been obtained from Kaggle (URL:

https://www.kaggle.com/datasets/gauravsrivastav2507/ehr-dataset?resource=download).

This dataset contains information about

- the cancer patients,
- their tumors,
- the treatments they underwent, and
- their survival statuses.

However, the type of cancer has not been specified in the dataset. Also, how the data have been collected, is not mentioned in the aforesaid website.

1.4. Brief Literature Survey

So far, many scientists have worked on cancer patient survival prediction, using different datasets. As for example, Kang et al. tested the efficiency of some Machine Learning models in predicting mortality among breast cancer patients, and found the Extra Survival Trees (EST) model to be the best among them [2], Vale-Silva and Rohr noted that when MultiSurv, a multimodal Deep Learning method, was applied to data corresponding to 33 different types of cancer, it generated accurate pan-cancer patient survival curves [3], Nunez et al. examined the effectiveness of some Natural Language Processing (NLP) models for predicting cancer patient survival, and observed that all the models performed

well [4], Hao et al. employed a Deep Learning technique for cancer survival prediction, and noticed that it performed better than three efficient Deep Learning models (used for the prediction of cancer patient survival) [5]. However, no such work has been found where the dataset used in this project, has been utilized. A cancer patient survival study with a hitherto unused dataset and/or method can show the effectiveness of the pertinent dataset and/or method for predicting the probability of survival of cancer patients.

2. Methodology

A brief discussion on the data and the method, used in this work, is given below.

2.1. Data Description

Dataset, used here, has been obtained as a .csv file. It contains 20,000 rows and 9 columns. The columns of the dataset are:

- Patient_ID: alphanumeric data,
- Age: integral data,
- Gender: alphabetic data,
- Tumor_Size(cm): floating-point data,
- Tumor_Type: alphabetic data,
- Biopsy_Result: alphabetic data,
- Treatment: alphabetic data,
- Response_to_Treatment: alphabetic data, and
- Survival_Status: alphabetic data.

Fig.-2 shows how the dataset, utilized here, looks like.

Patient_ID	Age	Gender	Tumor_Size(cm)	Tumor_Type	Biopsy_Result	Treatment	Response_to_Treatment	Survival_Status
c044501a-43ca-4a0c-8b8b-991439ba1b6a	52	Female	5.08	Benign	Positive	Surgery	No Response	Survived
b8900c4c-1232-4084-9432-5d02eba74d20	32	Female	0.8	Benign	Negative	Surgery	Complete Response	Survived
3004e2bc-8037-49cb-a542-d5612b73beab	70	Female	9.56	Benign	Positive	Radiation Therapy	Complete Response	Deceased
1df86af7-6745-4dea-b127-cbc9915079fc	21	Female	3.07	Malignant	Negative	Surgery	Partial Response	Survived
128e00c3-72e3-4031-a7f4-1165d7199cce	62	Male	7.17	Malignant	Positive	Radiation Therapy	Complete Response	Deceased
2b3cc8d5-f2f7-4ce2-be51-f5f489b53244	60	Female	8.31	Benign	Negative	Radiation Therapy	Complete Response	Deceased
2f8c5926-bedb-418e-84ed-8b05412e495f	34	Female	0.66	Benign	Negative	Chemotherapy	Complete Response	Survived
315f3ae6-b44c-42a8-9ce3-deb728a662f9	69	Female	2.2	Malignant	Negative	Surgery	Partial Response	Survived
494e30bf-b2e5-46b1-8323-ec4b1ef6ff1b	49	Male	6.42	Benign	Negative	Chemotherapy	No Response	Deceased
b55064cc-82e5-4d22-801a-32c2a9106a44	80	Female	2.04	Malignant	Negative	Chemotherapy	Partial Response	Deceased
f9763e16-ab67-4206-9468-65384c167f39	59	Female	5.75	Malignant	Negative	Radiation Therapy	No Response	Survived
4e9ae3b3-f55c-409e-8d81-79051572acf1	55	Female	0.81	Benign	Positive	Surgery	No Response	Deceased
20f857d7-7485-4056-bd9f-5dff393880ff	44	Male	10	Malignant	Positive	Chemotherapy	Complete Response	Deceased
48a387ea-725a-4071-8e9a-61f5f09cbaa6	54	Female	0.63	Malignant	Positive	Surgery	Partial Response	Survived
5a2de590-520f-40a5-ade6-6383bea4f85f	29	Male	3.46	Malignant	Negative	Radiation Therapy	Partial Response	Deceased
af968cb7-b465-40a4-b64f-cc85cc1b9e68	65	Male	8.7	Benign	Negative	Radiation Therapy	No Response	Survived
a5283707-91c4-416f-ab72-a39a903b877f	56	Female	4.37	Malignant	Positive	Surgery	No Response	Survived
5a78979c-891e-4380-800d-99ec7a6a749e	30	Female	3.43	Malignant	Negative	Chemotherapy	Complete Response	Survived
49ff41f7-3423-49d4-83d1-4779342dc2db	74	Male	1.01	Benign	Positive	Chemotherapy	No Response	Deceased
c7293d23-0f41-4d60-b3e2-d378758e28c9	23	Female	8.13	Malignant	Positive	Chemotherapy	Partial Response	Deceased

Figure 2 A small part of the dataset used in this study

2.2. Machine Learning Algorithm

In this work, the following two widely used Machine Learning algorithms have been employed:

- Logistic Regression
- Support Vector Machine

2.2.1. Logistic Regression

Logistic Regression is generally employed to evaluate the probability that an instance belongs to a particular class (e.g., the probability that for a particular patient (an instance), the 'survival status' is 'survived' (a class)). If the probability is greater than 50%, Logistic Regression model predicts that the instance belongs to the relevant class (e.g., if the probability is greater than 50%, the 'survival status' of the particular patient is 'survived'). Otherwise, it predicts that the instance does not belong to that class, which implies that the instance belongs to the other class (e.g., if the probability is not greater than 50%, the 'survival status' of the particular patient is not 'survived'; rather, it is 'deceased' (the other class)).

Expression for the probability (p) is given by equation-1:

$$p = \sigma (\theta^T . \mathbf{x}) \quad \text{.....(1)}$$

where,
σ = a sigmoid function,
θ = the parameter vector of the Linear Regression model (the model uses these parameters to calculate a weighted sum of input features),
θ^T = transpose of θ, and
x = the input feature vector of a particular instance (examples of input feature: 'Age', 'Gender', 'Tumor_Type', 'Treatment' etc.).

2.2.2. Support Vector Machine

Support Vector Machine tries to create a boundary with the largest possible margin between two classes (viz., 'survived' and 'deceased'), so that these classes can be easily separated. If the classes are easily separable, the instances can be unambiguously categorized into these two classes. Often, a dataset is not linearly separable between two classes. One approach to tackle the above problem is to add more features (viz., polynomial features), so that the dataset can become linearly separable; another approach is to add features computed with the help of a similarity function (which indicates each instance's resemblance with a particular landmark), to make the dataset linearly separable (suitable instances may be selected as landmarks).

2.3. Method

The method employed in this project, involves the following steps:

- Dataset has been checked to find out the presence of any null value in it. No null value has been detected.
- 'Patient_ID' column has been deleted; this is because, the patient id will be of no use in predicting the 'Survival_Status' of the patients.
- One hot encoding technique has been applied to the columns of 'Gender', 'Tumor_Type', 'Biopsy_Result', 'Treatment' and 'Response_to_Treatment', and label encoding has been applied to 'Survival_Status', to convert the alphabetic data of these columns into numeric (binary) format. ['Gender' column is replaced by 'Male_Pt' column and 'Female_Pt' column; 'Tumor_Type' is replaced by 'Benign_Tmr' and 'Malignant_Tmr'; 'Biopsy_Result' is replaced by 'Pos_Biop_Rs' and 'Neg_Biop_Rs'; 'Treatment' is replaced by 'Srgry', 'Rad_Thrpy' and 'Chmothrpy'; 'Response_to_Treatment' is replaced by 'No_Rspns', 'Cmplt_Rspns' and 'Prtl_Rspns'; in case of 'Survival_Status', the 'Survived' class is assigned the value 0, and the 'Deceased' class is assigned the value 1].
- Min-max scaling has been applied to the columns of 'Age' and 'Tumor_Size(cm)', to bring the numeric data of these columns in the same range.
- The dataset has been divided in 80:20 ratio to create the training data (80%) and the test data (20%).
- Logistic Regression and Support Vector Machine have been applied with the default hyperparameters. These two models have been trained with the training dataset, and tested with the test dataset. The accuracies of their performances have been recorded.
- The Python programming language has been used to implement the above steps.

3. Results and Discussions

It has been found that the accuracy of prediction of the survival status of cancer patients is 51.6% for Logistic Regression, and 52.2% for Support Vector Machine.

The above result indicates that:

- The performance of neither model is commendable, and
- The support vector machine technique has performed slightly better than the logistic regression algorithm.

The possible reasons for this poor performance may be:

- The quality of the data is probably not very good,
- The data are probably insufficient for predicting the survival status, and
- The machine learning models, used here, are probably not capable of giving a better result for this dataset.

However, considering the fact that both Logistic Regression and Support Vector Machine are efficient and widely used models, the actual reason for the poor performance of these two algorithms seems to be the low quality and/or the insufficiency of the data.

4. Conclusion

This study has attempted to predict the survival status of cancer patients by employing two well-known Machine Learning techniques viz., Logistic Regression and Support Vector Machine, and using a dataset of Kaggle. However, neither of these models has performed satisfactorily, and the actual cause of this poor performance appears to be the low quality and/or the insufficiency of the data used.

If possible, Logistic Regression and Support Vector Machine may be applied, in future, on other similar datasets for predicting the survival status of cancer patients; also, other Machine Learning/Deep Learning models may be employed for this purpose.

Compliance with ethical standards

Acknowledgments

I am thankful to all the faculty members of Global Institute of Health Science, for providing me the opportunity to work on this project.

References

[1] National Cancer Institute at the National Institutes of Health, "What Is Cancer?," 2021. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[2] H. Y. J. Kang, M. Ko, and K. S. Ryu, "Prediction Model for Survival of Younger Patients with Breast Cancer Using the Breast Cancer Public Staging Database," Sci. Rep., vol. 14, 2024.

[3] L. A. Vale-Silva and K. Rohr, "Long-Term Cancer Survival Prediction Using Multimodal Deep Learning," Sci. Rep., vol. 11, 2021.

[4] J.-J. Nunez, B. Leung, C. Ho, A. T. Bates, and R. T. Ng, "Predicting the Survival of Patients with Cancer from Their Initial Oncology Consultation Document Using Natural Language Processing," JAMA Netw. Open, vol. 6, no. 2, 2023.

[5] Y. Hao, X.-Y. Jing, and Q. Sun, "Cancer Survival Prediction by Learning Comprehensive Deep Feature Representation for Multiple Types of Genetic Data," BMC Bioinformatics, vol. 24, 2023.

Author's Short Profile

Diptarshi Mitra (b. 1981) has an MTech degree in Remote Sensing and Geographical Information System (specialization: Geoinformatics) from Indian Institute of Remote Sensing, Dehradun (2018), and another MTech degree in Computer Science and Engineering (with specialization in Data Science) from JIS Institute of Advanced Studies and Research Kolkata (2023). He also has a Certificate in Health Informatics & Management from Global Institute of Health Science, Ahmedabad (2024). Currently, he is pursuing PhD in Machine Learning at JIS Institute of Advanced Studies and Research Kolkata.