EasyChair Preprint

№ 8856

Risk Prediction Analysis of Venous Thromboembolism

Likhitha Mannaru, Siva Prasad Reddy Kudumula Venkata, Manjusha Kambam and Sruthi Kurlu

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

September 22, 2022

RISK PREDICTION ANALYSIS OF VENOUS THROMBOEMBOLISM

M.Likhitha[1], K V Siva Prasad Reddy[2], K.Manjusha[3],K.Sruthi[4]

[1]Student,[2]Assistant Professor, [3,4]Student

CSE Department, JNTUACEP, Pulivendula, AP ,India.

[1] mannarulikhitha@gmail.com,[2]sivajntuacep@gmail.com,[3]kambammanjusha@gmail.com,
[4]kurlusruthireddy73@gmail.com.

ABSTRACT

venous Thromboembolism(VTE) is an illness that happens once a blood clot occurs in veins. It is necessary to grasp concerning Deep Vein Thrombosis as a result that occurs to the humans and may occurs severe disability and sometimes it is fatal. Medical records obtained are not balanced, that is, it include a lot of patients who failed to endure thrombotic cases than patients who tough them what will cause False predictions, particularly for some patient cluster at high risk of occlusion. By this event of medical prognostic models, personalised models may occur, serving to the practician to realizes much stability among danger of trauma, therefore the threat leads to thrombosis.

KEYWORDS

Venous thromboembolism, Deep vein thrombosis, multi-objective improvement, antithrombotic prophylaxis.

1.INTRODUCTION

Pulmonary thromboembolism associated with thrombotic events, namely deep vein thrombosis, remains the cause of morbidity associated with thrombosis. related to cancer [1]. These events are associated with neoplastic hypercoagulability, chemotherapy, vascular injury from invasive procedures (eg, central venous catheterization), and prolonged postoperative rest [2]. Since Trousseau's first report on the association between thromboembolism and cancer, physicians are worried with thrombotic problems and in urge for early prevention and surgery [3], [4].

One of the greatest threat of Venous thromboembolism of solid tumours is connected to epithelial ovarian cancer [5], [6]. Moreover, retrospective studies in patients with ovarian cancer denotes as the endurance maybe high-flown with the existance of Venous thromboembolism [7]. Patients with cancer are considerably a lot of possible to develop VTE than folks while not cancer and skill higher rates of VTE return and injury complications throughout VTE treatment. Within the nineteenth century, physicians are involved concerning thrombotic events and also the want for prophylaxis and prior medical care. When a part of the clot is interrupted it may travel along the blood flow to the lungs, leads to occlusion called as pulmonary embolism which is a serious problem of DVT. If the clot is small, people can recover from PE with right treatment. But, this might be harmful to lungs. If blood clot is high, it block haemoglobin to reach lungs, can cause death. Members having this problem will experience pain, swelling, scaling or ulcers. situations at which, this manifestation might be very serious.

There may be a clinical want for strong models in detecting, that victims are at danger. Therefore, need antithrombotic prophylaxis. When antithrombotics are suggested to the victims,attentive tracking is implemented, Speedy motion is suggested when hemorrhage occurs. But, the recommendation of antithrombotics for the outpatients is harmful due to the fact the control methods to be had in a sanatorium aren't to be had outside. Bearing in thoughts that there aren't any ideal fashions, the mistake prices required for an outpatient rating can be unique than the ones required for an inpatient rating. For outpatients, false

positives are decreased in order that antithrombotics aren't prescribed to sufferers who do now no longer want them, thereby growing their danger of bleeding. However, for hospitalized sufferers, the focal point have to be on false negatives (the ones sufferers who have been categorized as less threat and have been cast aside with out prophylactic remedy although they have been in excessive danger). The need to save you thrombotic activities in sufferers hospitalized after surgical procedure via way of recommending antithrombotic prophylaxis.

2.LITERATURE REVIEW

In this paper we implemented the theoretical part of risk prediction. There is one tool which each and every information scientist ought to use or should be snug with, it's Jupyter Notebooks. Jupyter Notebooks are powerful, versatile, shareable and supply the flexibility to perform data visualisation within the same environment. Jupyter notebooks will illustrate the analysis method step by step by composing the things like code, images, text, output etcetera in a very step by step manner.

3.RESEARCH METHODOLOGY

3.1.DATA:

Patients were hospitalized and medical data was retrieved from electronic medical data of venous thromboembolism.

The clinical data from the electronic medical records of venous thromboembolism were collected from the patients who were admitted to the hospital

> The pertinent medical reports of each patient includes age, BMI(kg/m^2), Haemoglobin(g/dl), Red Blood Cells count(10^6/ml), White Blood Cells count(10^9/l), Haematrocrit(%), platelets(10^-3/ml),neutrophils(%),lymphocytes(%),absolute neutrophil count(10^-3/ml),CA125(U/ml),MCV(fl),MCHC(g/dl),Reticulocytes(%),CRP(mg/l),Fibrinogen(mg/dl),TNF.

3.2. FEATURE SUBSET SELECTION

To avoid overfitting the model, a ratio of at least 10:1 should be used among the number of patients and predictors [8], [10]. Feature selection is a manner of choosing the subset of the maximum applicable capabilities from the authentic features set through putting off the redundant, beside the point, or noisy features. While growing the machine learning model, only some variables withinside the dataset are beneficial for constructing the model, and the rest features are both redundant or beside the point. If we enter the dataset with some of these redundant and beside the point capabilities, it can negatively effect and decrease the general overall performance and accuracy of the model.

The 17 feasible predictors to be had for the sufferers will be used so we did now no longer set up a most wide variety of capabilities to be decided on and we did now no longer examine prototype with bigger wide variety of capabilities. But, easier prototypes were selected for many causes besides controlling the dimensionality problem, such as removing extraneous and other, so developing the analysis performance type [11].

3.3. DATA PREPROCESSING

Data pre-processing is the first step that marks the beginning of the process. Often, real-world data is inconsistent, inaccurate, incomplete, and sometimes missing specific attribute values/trends. this can happen anywhere data preprocessing is done - it helps to clean, format and organize the information, making it ready for machine learning models. Data preprocessing in machine learning refers to the technique of preparing raw data to make it suitable for building and framing machine learning models. Simply put, data preprocessing in Machine Learning can be a data processing technique that turns raw data into an understandable and readable format.

3.4. PERFORMANCE METRICS

Performance metrics are measurements that should be increased else decreased when building a classifier model and allow classifier performance to be measured and reported. There are a number of metrics and their correct selection is necessary to make further decisions [13]. Accuracy is the primary metric that indicates how often a classifier detects correctly. This live isn't a helpful metric within the case of unbalanced databases as a result of it primarily indicates achievement of bulk class. Sensitivity determines how often a classifier correctly detects a Venous thromboembolism prediction for hazardous threat patients. Opposite hand, exactness expounded to frequency of which patients expected to be danger can actually leads to VTE whether it hardly cured. But, only the accuracy is flawed as that does not provide awareness to diversity that is illustrated from hazardous threat victims specifically illegitimate further more with the minimal threat. Especially likelihood of victims with DVT meant to be considered at high risk.

TN	FP
FN	TP

3.5. CLASSIFICATION MODEL

Although several kinds of Machine Learning models are there that can perform this categorization, but choosing best method relays upon many factors, including kind-of data set, the calculation time and implementation of outcomes [9].

Here, we performed SVM algorithm. Support Vector Machine (SVM) is a supervised machine learning algo rithmic program used for each classification. Although we are saying regression issues yet its best suited to c lassification. the target of SVM algorithm is to search out a hyperplane in an N-dimensional area that clearly classifies the info points. The dimension of the hyperplane is based on the quantity of options. If the number of input features is two, then the hyperplane is simply a line. If it is three, then the hyperplane becomes a 2-D plane. It becomes tough to imagine once the number of features exceeds three. The accuracy obtained for t his classification algorithm is 0.909091 with confusion matrix

Array ([[4, 1],
[0, 6]], dtype=int64)

Also in this research paper we performed Logistic regression. Logistic regression is generally a supervised t ype prototype. The prototype builds a regression version to are expecting the possibility that a given informa tion access belongs to the class numbered as “1”. Similarly Linear regression assumes that the information f ollows a linear function, Logistic regression fashions the information the usage of the sigmoid function. Log istic regression will become a classification method best whilst a selection threshold is introduced into the pi cture. The putting of the threshold cost is a totally critical factor of Logistic regression and is depending on t he classification problem itself. The accuracy obtained in this Logistic regression is 0.909091 which is simil ar to SVM with confusion matrix

Array ([[4, 1],
[0, 6]], dtype=int64)

We also performed k-nearest neighbour algorithm. k Nearest Neighbors is a kind of classification where the function is merely approximated locally and every calculation is postponed till evaluation. As this prototype depends on distance for categorization, if options indicates totally different scales,Then normalizing the training knowledge will increase its accuracy. Each for classification and regression, a helpful technique is to assign weights to the contributions of the neighbors, in order that the nearer neighbors contribute more to the average than the more distant ones. The accuracy obtained is 0.72723 with confusion matrix

Array ([[3, 2],
[1, 5]], dtype=int64)

4.RESULTS

We first tested the logistic regression with patient data from this study, getting the results shown in the table. The data reviewed in this paper, the logistic regression classifier showed an accuracy of 90%. And we did Knn with the data from our study, the knn classifier showed an accuracy of 72.7%. showed that the accuracy similar to logistic regression was 90%. Confusion matrix, reflecting that hazardous victims are the patients experienced to DVT and median threat is calculated to be free of this event based on results

CLASSIFICATION REPORT	*PRECISION*	*RECALL*	*F1-SCORE*	*SUPPORT*
0	*1.00*	*0.80*	*0.89*	*5*
1	*0.86*	*1.00*	*0.92*	*6*
accuracy			*0.91*	*11*
Macro avg	*0.93*	*0.90*	*0.91*	*11*
Weighted avg	*0.92*	*0.91*	*0.91*	*11*

4.1.Classification report for logistic regression

CLASSIFICATION REPORT	*PRECISION*	*RECALL*	*F1-SCORE*	*SUPPORT*
0	*0.75*	*0.60*	*0.67*	*5*
1	*0.71*	*0.83*	*0.77*	*6*
accuracy			*0.73*	*11*
Macro avg	*0.73*	*0.72*	*0.72*	*11*
Weighted avg	*0.73*	*0.73*	*0.72*	*11*

4.2.Classification report for k-nearest neighbour

CLASSIFICATION REPORT	*PRECISION*	*RECALL*	*F1-SCORE*	*SUPPORT*
0	*1.00*	*0.80*	*0.89*	*5*
1	*0.86*	*1.00*	*0.92*	*6*
accuracy			*0.91*	*11*
Macro avg	*0.93*	*0.90*	*0.91*	*11*
Weighted avg	*0.92*	*0.91*	*0.91*	*11*

4.3.Classification report for SVM

5.CONCLUSION

Here, Multi-objective ML-based clinical models are predicted, even though same methods have been implemented prior[12]. This method has never been utilized before for the prediction of DVT in ovarian cancer victims, managing medical imbalances. If a physician intends to use an extremely aggressive prophylactic antithrombotic regimen, he or she may choose to use a model with an infrequent false-positive probability, for prophylactic treatment of patients without thrombosis is low. In turn, this increases the

trouble of false negative.so, a significant amount of victims will develop this disease as they don't need to benefit from antithrombotic prophylaxis. On the other side, we will look at a reliable prophylactic surgery that doctors prefer for a better true positive rate. However, it can also increase the false-positive rate, and thus more patients may receive inappropriate prophylaxis, but thus thrombosis will be eluded for a higher amount of patients.

6.REFERENCES

[1] M. Dutia, R. H. White, and T. Wun, "Risk assessment models for cancer-associated venous thromboembolism," Cancer, vol. 118, no. 14, pp. 3468–3476, 2012.

[2] M. Dicato, "Venous Thromboembolic Events and ErythropoiesisStimulating Agents: An Update," Oncologist, vol. 13, no. Supplement 3, pp. 11–15, 2008.

[3] V. De Stefano, "Arterial thrombosis and cancer: the neglected side of the coin of Trousseau syndrome," Haematologica, vol. 103, no. 9, pp. 1419–1421, 2018.

[4] G. H. Lyman et al., "Venous Thromboembolism Prophylaxis and Treatment in Patients With Cancer : American Society of Clinical Oncology Clinical Practice Guideline Update 2014," J. Clin. Oncol., vol. 33, no. 6, pp. 654–656, 2015.

[5] A. A. Khorana, N. M. Kuderer, E. Culakova, G. H. Lyman, and C. W. Francis, "Development and validation of a predictive model for chemotherapy-associated thrombosis," Blood, vol. 111, no. 10, pp. 4902–4907, May 2008.

[6] B. Bhagya Rao, R. Kalayarasan, V. Kate, and N. Ananthakrishnan, "Venous Thromboembolism in Cancer Patients Undergoing Major Abdominal Surgery: Prevention and Management," ISRN Vasc. Med., vol. 2012, pp. 1–22, 2012.

[7] C. Fotopoulou, A. N. Karavas, R. Trappe, and B. Aminossadati, "Incidence of Venous Thromboembolism in Patients With Ovarian Cancer Undergoing Platinum / Paclitaxel – Containing First-Line Chemotherapy : An Exploratory Analysis by the Arbeitsgemeinschaft Gynaekologische Onkologie Ovarian Cancer Study Group," J. Clin. Oncol., vol. 26, no. 16, pp. 2683–2689, 2008.

[8] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis.," Cancer Inform., vol. 2, pp. 59– 77, Feb. 2007.

[9] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Comput. Struct. Biotechnol. J., vol. 13, pp. 8–17, 2015.

[10] D. Chicco, "Ten quick tips for machine learning in computational biology," BioData Min., vol. 10, no. 1, pp. 1–17, 2017.

[11] P. Refaeilzadeh, L. Tang, and H. Liu, "On comparison of feature selection algorithms," in Proceedings of AAAI workshop on evaluation methods for machine learning II, 2007, vol. 3, no. 4, p. 5.

[12] Y. Jin, S. Member, B. Sendhoff, and S. Member, "Pareto-Based Multiobjective Machine Learning : An Overview and Case Studies," vol. 38, no. 3, pp. 397–415, 2008.

[13] M. Hossin and M. N. Sulaiman, "A Review on Evaluation Metrics For Data Classification Evaluations," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 1–11, 2015.