ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.6, 3487–3495
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.6.28 Spec. Issue on RDESTM-2020

BIOMEDICAL DATA ANALYSIS IN PREDICTING AND IDENTIFICATION CANCER DISEASE USING DUO-MINING

NARASIMHACHARY CHOLLETI[1] AND TRYAMBAK HIRWARKAR

ABSTRACT. As an option in contrast to the dull physical stockpiling of assets it is significant build up a data stockroom explicit to cancer disease and a data mining model to anticipate cancer prior. On the off chance that an AI strategy is created to store an individual's medical and general record and anticipate his inclination towards cancer, its sort and precise symptomatic technique, doctors can legitimately begin treatment quickly without burning through the valuable time in various strategies for diagnosis [1]. There have been numerous data mining procedures in health care and partnered ventures and explicitly regarding single sort of cancer. This examination centers around the structure of multidimensional cancer data stockroom and advancement of data mining model for the early identification of six sorts of cancer, henceforth counteractive action is additionally conceivable [2].

1. INTRODUCTION

Cancer, also referred to as "risk," is a social event characterized by situations that require the strange progression of cells at risk of aggression or spread to various body elements. Not all tumors are cancerous; Genital tumors do not spread to different parts of the body. Possible signs and symptoms include: other abnormalities, abnormal conduction, delayed hacking, unexplained weight loss, and a change in craps. Although these reactions can trigger cancer, they can also

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 62P10.

Key words and phrases. Duo mining, SVM, ANN, cancer.

be caused by various problems [3]. There are more than 100 known cancers that affect individuals. Smoking is the cause of approximately 22% of cancer cases. A 10% weight gain is due to an unpleasant diet, physical activity and alcohol consumption. The various components dissolve the input of certain pollutants, ionizing radiation and regular toxins. Approximately 20% of cancers are directly caused by stains such as hepatitis B, hepatitis C and human papilloma virus (HPV) contamination [4-5]. These factors almost always manifest themselves by changing the properties of a cell. As a rule, such genetic changes are necessary before cancer develops. Approximately 5 to 10% of cancers are the direct result of hereditary malformations of human origin. Cancer can be identified by clear signs and reactions or screening tests. It is then usually deepened by therapeutic imaging and detected by biopsy.

Do not smoke, maintain high weight, do not use alcohol, eat too much vegetables, eat cereals, get yourself vaccinated against some dangerous diseases, do not eat a ton of red meat, do not expose to plenty of light. Screening is useful for cervical cancer and colorectal cancer. The benefits of breast cancer prevention are controversial. Cancer is regularly treated with a mix of radiotherapy, medical intervention, chemotherapy and treatment. Pain and Manifestation Counseling is an important issue. Palliative attention is especially important in people with advanced disease. The likelihood of endurance depends on the type of cancer and the degree of disease at the beginning of the treatment. For young people under the age of 15 at the time of diagnosis, the global multi-year endurance rate is 80%.

1.1. **Data Mining.** Data Mining can be defined as a way to effectively recognize cloud models and instances in databases and to use this data to create intelligent models. On the other hand, it can be defined as the insurance and data validation system and the creation of models that use huge databases to explore past and dark models. Data Mining is a logical way of searching for large amounts of data, finding reliable models and effective connections between components, and then encouraging revelations by applying perceived guidelines to new subgroups.

Data mining has become a logical standard. It was determined that some segments were asked to use data mining applications. Data mining meetings can affect costs, salaries, and profitability while maintaining a remarkable level

of care. Data mining associations are better organized to meet their long-term needs. The data can be a striking field for state restorative associations, but they must first be converted into data. The use of data mining applications in the field of social protection is even more mobile. Data mining can provide vital data for all meetings with the food service industry.

2. Research Method

2.1. **Data Collection.** We gathered cancer patient informationfrom various hospitals in Tamilnadu. All patients of this database are men and women of age between 0-20 Years. The variable takes the value 'TRUE' and 'FALSE', where 'TRUE' means a positive test for T2DM and 'FALSE' means a negative test for T2DM.It is essential to look at the information with preprocessing which comprise of cleaning, change andintegration [7]. The analysisclinical attributes: Gender, family history, age, Habitual Smoker, cancer level.

2.2. **Methodology.** Duo-mining is likewise called as intelligent text analysis, text data mining in the text reveals beforehand imperceptible examples in existing assets. Duo mining includes the utilization of strategies from areas, for example, information retrieval, information extraction and data mining Text Mining itself isn't capacity, it joins various functionalities Information Extraction (IE), Searching, Summarization, Categorization, Clustering, Prioritization, Information Monitor and Information Retrieval [8]. Information Retrieval (IR) frameworks recognize the reports in an assortment that coordinate a client's inquiry. The significant strides in text mining were 1) Text Gathering and 2) Text Preprocessing. Text gathering incorporates an assortment of raw documents like Patient information which were in the text or script position and these records may contain unstructured data. The preprocessing stage begins with tokenization. Tokenization is a division of a report into terms. This procedure likewise alluded to as feature generation. In-text preprocessing the raw data as text records are gathered from text scripts or Flat files. The data is changed over into an organized arrangement and put away in Microsoft Access Database.

To do the Search and Information Extraction, we utilized a tool call Duo-Mining developed in Java as the part of the Text Mining Tool and converts the Un-structured data into the structures manner.

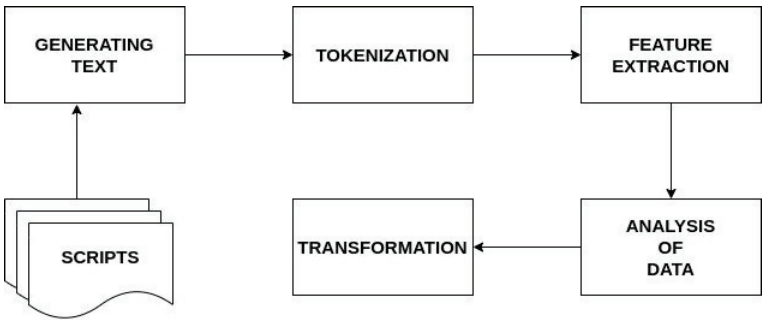

FIGURE 1. Conversion of Unstructured data into Structured Data Set using Duo-Mining Tool

3. PREDICTION MODELS

We utilized three distinct sorts of grouping models: ANNs, SVM, and strategic relapse. These models were chosen for incorporation in this study because of their notoriety in the recently published literature as well as the desirable performance they had shown in our preliminary comparative studies. What follows is a concise description of these three model.

3.1. **Logistic regression (LR).** Logistic regression (LR) expands the strategies for different regression examinations to consider conditions in which the end variable is absolute. In a framework, conditions including explicit results are very conventional. In the setting of evaluating an educational program, for instance, forecasts might be gotten for the dichotomous results of accomplishment/disappointment or progressed/not progressed. Similarly, in a clinical system, a result may be the presence/absence of disease.

3.2. **Artificial Neural Network (ANN).** Artificial Neural Network (ANN) is a data preparing model that is motivated by the procedure of organic anxious tasks, for example, the mind, strategy learning. The basic detail of this model is one of a kind structure of the data preparing framework. It is made of a lot of profoundly interconnected preparing factors (neurons) working as one to answer explicit issues. ANNs, similar to individuals, an examination by model. An ANN is arranged for a specific use, for example, design recognizable proof or data investigation, through a learning strategy. Concentrating on natural techniques expects acclimations to the synaptic affiliations that exist between the neurons. Neural networks, with their remarkable capacity to get intrigued by

mind-boggling or loose data, can be applied to acquire designs and find slants that are too dark to even consider being seen by either people or other PC systems. A certified neural system can be thought of as an "authority" in the class of information it has been given to examining.

3.3. **Support Vector Machine (SVM).** Support Vector Machine (SVM) is one of the supervised classification models that is mostly applied in the area of the cancer determination. SVM gathers basic examples from all groups, these examples are called support vectors, and it circulates the classes by making a straight capacity that isolates. SVM can be applied to a guide between the information vector to a high dimensionality space is made applying SVM that intends to discover the differently reasonable hyperplane that isolates the data set into classes. The Linear classifier plans to expand the separation between the choice hyperplane and the closest data point, which is known as the constrained separation.

Given the preparation datasets of the structure $(x1, c1), (x2, c2), ..., (x_n, c_n)$ where c_i is either 1 ("yes") or 0 ("no"), a SVM finds the ideal isolating hyperplane with the biggest edge. Condition (3.1) and (3.2) speaks to the isolating hyperplanes on account of detachable datasets.

$$w.x_i + b \leq -1, for c_i = -1, \tag{3.1}$$

$$w.x_i + b \geq +1, for c_i = +1. \tag{3.2}$$

The problem is to minimize |w| subject to constraint (3.2). This is called constrained quadratic programming (QP)optimization problem represented by:

minimize $(1/2)||w||^2$

(subject to $c_i(w.x_ib) \geq 1$) Sequential minimal optimization (SMO) is one of efficient algorithm for training SVM.

4. Results

In this examination, the models were assessed based on the accuracy measures talked about above (classification accuracy, sensitivity, and particularity). The outcomes were accomplished utilizing ten times cross-approval for each model and depend on the normal outcomes acquired from the test data set (the tenth overlap) for each overlay. In contrast with the above investigations,

we found that the choice tree model accomplished a classification accuracy of 0.9000 with a sensitivity of 0.9188 and explicitness of 0.7375. The logistic regression model accomplished a classification accuracy of 0.8961 with a sensitivity of 0.9130 and explicitness of 0.7361. The ANN model accomplished a classification accuracy of 0.9107 with a sensitivity of 0.9310 and explicitness of 0.7383. Be that as it may, the SVM model played out the best of the four models by accomplishing an accuracy of 0.9285 with a sensitivity of 0.9423 and explicitness of 0.7572. A disarray network is a lattice portrayal of the classification results. In a two-class expectation issue, (for example, the one in this examination) the upper-left cell means the number of tests delegated true when they are true (for example true positives), and the lower right cell means the number of tests delegated false when they are false (for example true negatives).

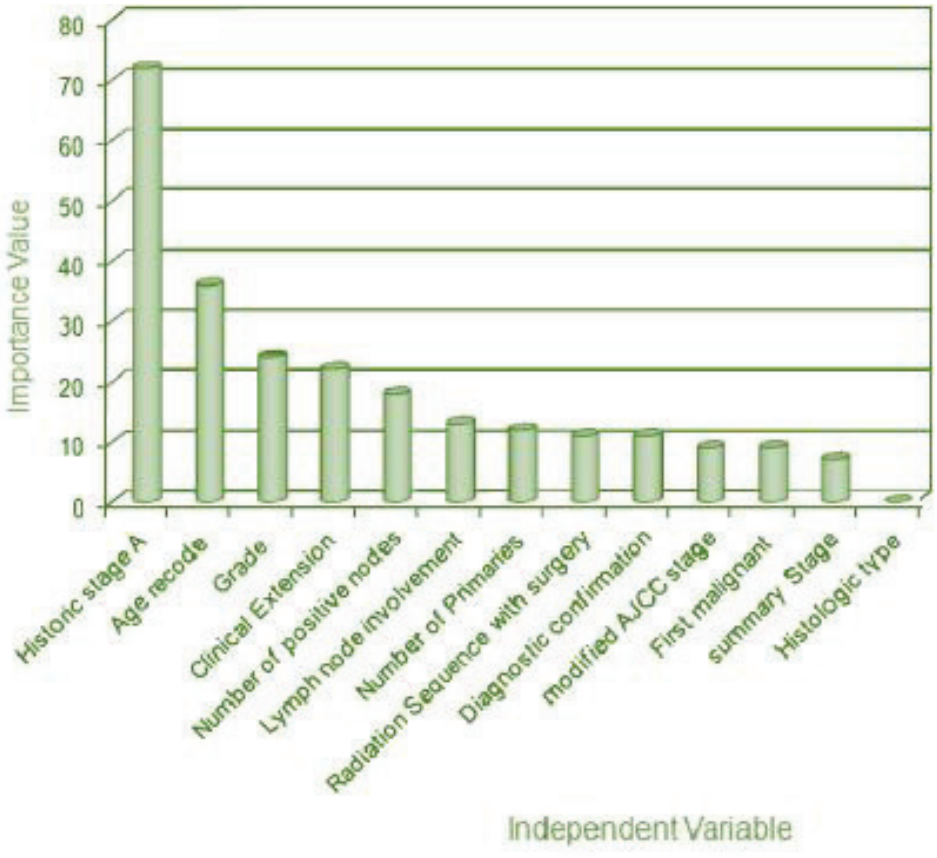

FIGURE 2. Variable importance chart for the ANN models

4.1. **Sensitivity analysis on ANN output.** We have utilized sensitivity analysis to increase some knowledge into the choice factors utilized for the classification issue. Sensitivity analysis is a strategy for extricating the circumstances and logical results connection between the data sources and yields of a neural system model. As has been noted by numerous examiners in the artificial knowledge field, more often than not ANNs may offer better prescient capacity however very little logical worth. This analysis is commonly true; in any case, sensitivity analysis can be performed to create understanding into the issue. As of late, it has become a usually utilized technique in ANN reads for distinguishing how much each information channel (free factors or choice factors) adds to the ID of each yield channel (subordinate factors). The sensitivity analysis gives information about the overall significance of the info factors in foreseeing the yield field(s). During the time spent performing sensitivity analysis, the ANN learning is crippled with the goal that the system loads are not influenced. The essential thought is that the contributions to the system are irritated somewhat, and the comparing change in the yield is accounted for as a rate change in the yield (Principe et al., 2001). The primary information is differed between its mean in addition to (or less) a client characterized number of standard deviations, while every single other information are fixed at their individual methods. The system yield is processed and recorded as the rate change above and underneath the mean of that yield channel. This procedure is rehashed for each info variable. As a result of this procedure, a report (typically a segment plot) is created, which outlines the variety of each yield as for the variety in each info. The sensitivity analysis performed for this examination venture and introduced in a graphical arrangement in Figure 2 records the information factors by their relative significance (from generally critical to least significant). The worth appeared for each information variable is a proportion of its relative significance among different factors.

5. CONCLUSION

Duo-mining has been extensively applied in the medical field and has filled in as a significant diagnostic tool that helps doctors in analyzing the accessible data just as planning a clinical master framework. In this paper, three-system were utilized to be specific SVM, Logistic Regression, ANN. Their philosophies

and primary features were depicted. The tenfold cross-validation approach was utilized in model structure and assessment, where we isolated the data set into 10 totally unrelated parcels utilizing a delineated inspecting method. Out of 10 folds, nine folds were utilized for preparing and the tenth fold was utilized for testing. We repeated this process 10 times so that each data point would be used as part of the training and testing data sets. For all tenfold accuracy, sensitivity, and explicitness were determined for every one of the three model sorts. At that point across tenfold accuracy, sensitivity, and explicitness were found the middle value of so as to make a correlation on the best model foreseeing prostate cancer survivability. By averaging across 10 folds we found that SVM models played out the best with an accuracy proportion of 92.85%. The ANN models approached second followed by the logistic regression models.

Other relevant references are [6, 9-15].

REFERENCES

[1] A. TOLOIEESHLAGHY, A. POOREBRAHIMI, M. EBRAHIMI, A. R. RAZAVI, L. GHASEM AHMAD: *Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence*, 2013.

[2] Y. LI, H. CHEN, L. CAO, J. MA: *A Survey of Computer-aided Detection of Breast Cancer with Mammography*, 2016.

[3] H. L. CHEN, B. YANG, J. LIU, D. Y. LIU: *A support vector machine classifier with rough set based feature selection for breast cancer diagnosis*, Expert Syst. Appl,. **38**(7) (2011), 9014–9022.

[4] I. KONONENKO: *Machine learning for medical diagnosis: history, state of the art and perspective*, Artif. Intell. Med., **23**(1) (2001), 89–109.

[5] V. CHAURASIA, S. PAL: *Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability*, 2017.

[6] J. R. QUINLAN: *Improved Use of Continuous Attributes in C4.5*, J. Artif. Intell. Res., **4** (1996), 77–90.

[7] H. J. HAMILTON, N. SHAN, N. CERCONE: *RIAC: A Rule Induction Algorithm Based on Approximate Classiication*, 1996.

[8] S. ARUNA, P. RAJAGOPALAN, L. V. NANDAKISHORE : *Knowledge based analysis of various statistical tools indetecting breast cancer*, Int.j. Stat., **5** (2011), 37–45.

[9] S. BASHIR, U. QAMAR, F. H. KHAN: *Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble*, Qual. Quant., **49**(5) (2015), 2061–2076.

[10] D. BAZAZEH, R. SHUBAIR: *Comparative study of machine learning algorithms for breast cancer detection and diagnosis*, in 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016, 1–4.

[11] R. SETIONO: *Generating concise and accurate classification rules for breast cancer diagnosis*, Artif. Intell. Med., **18**(3) (2000), 205–219.

[12] G. WILLIAMS: *Descriptive and Predictive Analytics*, Data Mining with Rattle and R, New York, NY: Springer New York, 2011, 171–177.

[13] S. KESAVAN, E. SARAVANA KUMAR, A. KUMAR, K. VENGATESAN: *An investigation on adaptive HTTP media streaming Quality-of-Experience (QoE) and agility using cloud media services*, Taylor and Francis, International Journal of Computers and Applications, **5**(5) (2016), 33–44.

[14] K. VENGATESAN, S. MAHAJAN, P. SANJEEVIKUMAR, S. MOIN: *The Performance Enhancement of Statistically Significant Bicluster Using Analysis of Variance*, Advances in Systems, Control and Automation, Lecture Notes in Electrical Engineering 442, Chapter No. 64.

[15] M. SANTOSH, A. SHARMA: *A Proposed Framework for Emotion Recognition Using Canberra Distance Classifier*, J. Comput. Theor.Nanosci., **16**(9) (2019) 3778–3782.

DEPT. OF COMPUTER SCIENCE AND ENGINEERING
SRI SATYA SAI UNIVERSITY OF TECHNOLOGY AND MEDICAL SCIENCES, SEHORE
BHOPAL-INDORE ROAD, MADHYA PRADESH, INDIA

DEPT. OF COMPUTER SCIENCE AND ENGINEERING
SRI SATYA SAI UNIVERSITY OF TECHNOLOGY AND MEDICAL SCIENCES, SEHORE
BHOPAL-INDORE ROAD, MADHYA PRADESH, INDIA