International Journal of Technology, Management and Social Sciences (IJTMSS)
Volume 1_Issue 1 [October-December 2022]
www.ijtmss.org

CATEGORY DETECTION OF CANCER PATIENTS AND ROLE OF MACHINE LEARNING AS AN ASTROLOGER

Karismita Medhi[1] and Bapan Kalita[2]

[1,2]Department of Mathematics, The Assam Royal Global University, Assam, India

Abstract

The fatal cancer develops to any age, sex for many reasons like smoking, use of alcohol, chewing tobacco, etc. Looking at the data of the patients, it is now a days possible to predict the category of the patient. This was earlier possibly exercised by the astrologers. This can be replaced by machine learning. To do these, the machine needs to be trained with some of the random data of the entire dataset. Once the machine gets trained, it can predict the category of the remaining data. Here, we have gathered the data of lung cancer and Indian liver patient to predict whether the patient is having liver cancer or not, predict the gender of the patient as male or female or to detect any category of the patient etc. The program has been done in Python environment using Jupyter notebook.

Keywords: *Machine learning; Logistic regression; Supervised learning; Astrology; Character detection*

1. Introduction

Cancer is a disease that occurs when cells in our bodies grow at a greater pace than usual. These aberrant cells form a bulge or excrescence. Some cancers boost fast cell production, while others cause cells to flourish and divide at a dimmer rate. Most malignancies have four stages. The stage is influenced by a variety of factors, including the size and location of the excrescence.

1.1 Types of Cancer

There are several types of cancer. Among them are: Lung Cancer, Liver Cancer, Colon Cancer, Pancreatic Cancer, Kidney Cancer, Skin Cancer, Thyroid Cancer

1.2 Machine Learning

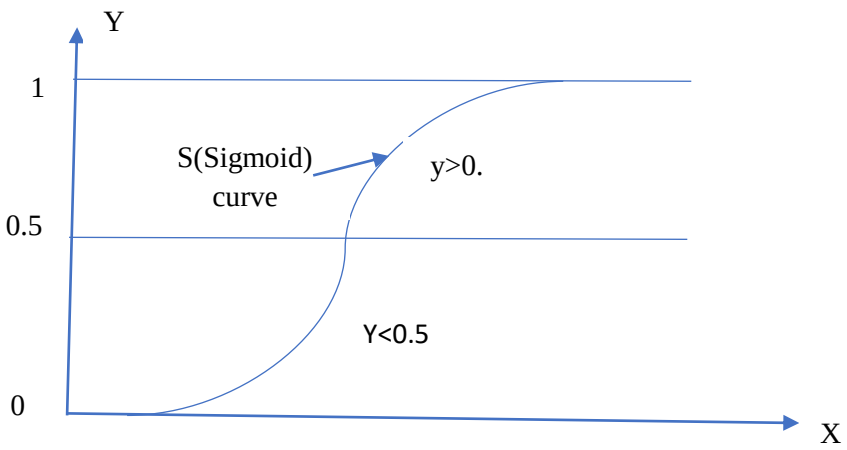
The phrase "machine learning" was coined in the 1950s by Arthur Samuel, an Artificial Intelligence colonist who built the first tone-literacy system for playing checkers. He discovered that the system worked better the further it was played. Machine Learning (ML) is a sort of artificial intelligence that enables software processes to predict

difficulties more accurately without being expressly designed to do so. To forecast new affair values, machine learning algorithms employ literal data as input. Machine learning is significant because it provides businesses with insights on trends in customer behaviour and company functional patterns, as well as assisting in the development of new products. Many of today's biggest organisations, such as Facebook, Google, and Uber, have made machine learning a core aspect of their operations.

1.3 Logistic Regression in Machine Learning

Logistic Regression is a common Machine Learning technique that falls under the Supervised literacy trend. It is used to forecast the categorical dependent variable using a collection of independent factors. A categorical dependent variable's affair is predicted using logistic regression. As a result, the offshoot must be a categorical or distinct value. It can be Yes or No, 0 or 1, True or False, and so on, but instead of reporting the precise value as 0 or 1, it presents the probabilistic values that fall between 0 and 1. Apart from how they are incorporated, Logistic Regression is extremely akin to Linear Regression. Logistic regression is used to solve bracket issues, whereas linear regression is used to solve regression problems. Rather of constructing a regression line, we fit a "S" shaped logistic function that predicts two maximum values in Logistic Regression (0 or 1). The logistic function curve represents the probability of commodities similar as whether the cells are malignant or not, if a mouse is overweight or not based on its weight, and so on. Logistic Regression is an important machine learning technique since it can offer opportunities and categorise new data using continuous and distinct datasets. It can classify compliances using various sources of data and quickly select the most effective factors for the bracket. The logistic function is seen in the graphic below:

Figure 1: Logistic Function



1.4 Logistic Function or Sigmoid Function

The sigmoid function is an excellent function for correlating predicted values to probabilities. It creates a map with any real value between 0 and 1 and another value. The logistic regression result must be between 0 and 1, and it cannot exceed this limit, forming a curve similar to the "S" shape. The Sigmoid function or logistic function is another name for the S-form curve. In logistic regression, we employ the notion of the threshold value, which specifies the likelihood of either 0 or 1. Values over the threshold value likely to be 1, while values below the threshold value tend to be 0.

1.5 Assumptions for Logistic Regression

a) The dependent variable must be of non-continuous or categorical kind.
b) There should be no multicollinearity in the independent variable.

1.6 Logistic Regression Equation

The logistic regression equation can be attained from the linear regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let us divide the above equation by

$$(1 - y); \frac{y}{1-y}; 0 \text{ for } y = 0, \text{ and infinity for } y = 1$$

- But we need range between –[infinity] to +[infinity], then take logarithm of the equation it will become:

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

1.7 Types of Logistic Regression

a) **Binomial Logistic Regression: O**nly two dependent variables of categorical kind such as 0 or 1, Pass or Fail, etc.
b) **Multinomial Logistic Regression:** 3 or more unordered categorical dependent variables such as “cat”, “dogs”, or “sheep”.

c) **Ordinal Logistic Regression:** 3 or more ordered categorical dependent variables such as "low", "medium", or "high".

1.8 Basic Mathematics of Logistic Regression

The basic mathematics of this regression model is:

$$\ln Odds(A) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

where,

$$\text{Odds (A)} = \frac{P(A)}{P(A')} = \frac{P(A)}{1-P(A)},\ 0 \le p \le 1,$$

we then define Odds (p) as

$$\text{Odds (p)} = \frac{p}{1-p}$$

For our purpose, Odds (p) can transform the probability function, which has values from 0 to 1, into an equivalent function with values between 0 and ∞. Taking natural log, we get a range of values from -∞ to ∞.

$$\text{logit (p)} = \ln \text{Odds (p)} = \ln \frac{p}{1-p} = \ln p - \ln (1- p)$$

$$\text{logit } (\pi) = \ln \text{Odds } (\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

where $\pi = P(A)$. it now follows that

$$\frac{P(A)}{1-P(A)} = \text{Odds (A)} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\varepsilon}$$

and so

$$p = P(A) = \frac{e^{b_0+b_1 x_1+b_2 x_2+\cdots+b x_k}}{1+e^{b_0+b_1 x_1+b_2 x_2+\cdots+b x_k}} = \frac{1}{1+e^{-(e^{b_0+b_1 x_1+b_2 x_2+\cdots+b x_k})}} = \frac{1}{1+e^{-b_0-\sum_{j=1}^{k} b_j x_j}}$$

Here, we shift to the model based on the observed sample (and so the parameter π is replaced by its sample estimate p, and β_j coefficients are replaced by the sample estimate b_j and the error term ε is dropped). For our purpose, we take the event A to be that the dependent variable y has a value of 1. If each y takes only the values 0 or 1, we can think of A as success and the complement A′ of A as failure. This is the concept for the trials of binomial distribution. For the regression model studied in Regression and Multiple Regression, a sample consists of n data elements of the form (y_i, x_{i1}, x, …, x_{jk}), but for logistic regression, each y_i only takes the value 0 or 1. Now let A_j = the event that $y_i = 1$ and $p_i = P(A_j)$. Similar to the regression line, this one predicts the

value of the dependent variable y based on the values of the independent variables $x_1, x_2, x_3, \dots, x_k$ and we get

$$p = P(y = 1) = \frac{1}{1+e^{-b_0-\sum_{j=1}^{k} b_j x_j}}$$

$$\text{logit}(p) = \ln \frac{p}{1-p} = \ln \frac{P(y=1)}{1-P(y=1)} = b_0 + \sum_{j=1}^{k} b_j x_j$$

and $\text{var}(y_i) = p_i(1-p_i)$.

For k = 1, we get

$$p = \frac{1}{1+e^{-b_0-b_1 x}}$$

1.9 Use of Logistic Regression

Sometimes the assumptions required for conducting multiple regression do not hold good. In that case, logistic regression can be used if:

a) The assumption of normality is not achieved.
b) The homoscedasticity is failed.
c) The predicted values will fall outside 0 and 1.

1.10 Odds Ratio

Definition 1: This is defined as follows:

$$R_{xi,x_j} = \frac{Odds(x_{i1}, \dots\dots, x_{jk})}{Odds(x_{jk} \dots\dots, x_{jk})}$$

$$= \frac{e^{b_0+\sum_{m=1}^{k} b_m x_{im}}}{e^{b_0+\sum_{m=1}^{k} b_m x_{Jm}}}$$

$$= e^{\Sigma_m^k} = 1^b m(x_i - x_j)$$

And $logit\ \frac{p_{x+1}}{p_x}$ using the notation $p_c = P(x)$.

1.11 Interpreting the odds ratio

In the case where,

$$log(p_{x+1} / p_x) = ln\left(\frac{p_{x+1}}{1-p_{x+1}} / \frac{p_x}{1-px}\right)$$

$$= \ln \frac{p_{x+1}}{1-p_{x+1}} - \ln \frac{p_x}{1-p_x}$$
$$= b_0 + b,(x+1) - b_0 - b_1 x = b_1$$

Thus,

$$\frac{Odds(x+1)}{Odds(x)} = \frac{p_{x+1}}{1-p_{x+1}} / \frac{p_x}{1-p_x} = e^{b_1}$$

Again, for any d

$$\frac{Odds(x+d)}{Odds(x)} = e^{b_1 d}$$

In case x behaves as a dichotomous variable,

$$e^{b_1} = \frac{Odds(1)}{Odds(0)}$$

1.12 Maximum Log-Likelihood

Since our model is based on binomial distribution, so the likelihood function is given by

$$L = \prod_{i=1}^{n} p_i^{yi} (1 - P_i)^{1-yi}$$

We derive the following definition by taking the natural log of both sides and simplifying it.

Definition 2: The log-likelihood function is:

$$LL = ln\, L = \sum_{i=1}^{n} [y_i\, ln\, p_i + (1 - y_i)\, ln(1 - p_i)]$$

where, the y_i are the observed values while the p_i are the corresponding theoretical values.

We find the maximum value of LL and this will enable to estimate b_i coordinates.

Astrologer: Astrologer is a typical person who can predict the future based on some available information.

2. Literature Review

To know about cancer and its detection through machine learning, we have studied some literatures which are narrated below:

Rahman et al. (2019): Chronic liver disease is the biggest cause of mortality worldwide, affecting a large number of people. This condition is caused by a variety of factors that affect the liver. This illness diagnosis is both expensive and complex. Their research aims to assess the efficacy of various machine learning algorithms in order to lower the high cost of chronic liver disease diagnosis through prediction. They employed six algorithms in their research: Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine, Nave Bayes, and Random Forest. The performance of several classification systems was assessed using various measuring techniques. They discovered 75%, 74%, 69%, 64%, 62%, and 53% accuracy for LR, RF, DT, SVM, KNN, and NB. According to analysis results, LR earned the maximum accuracy [1].

Nankani et al. (2019): Cancer is becoming one of the most common diseases that affects a large number of individuals. A molecular pathologist chooses a list of inheritable variants of interest to investigate. The molecular pathologist searches the medical literature for evidence that is relevant to the inheritable variants of interest. Finally, this molecular pathologist spends a significant amount of time discovering the evidence that is associated with each of the alterations in order to categorize them. In this research, they use machine learning methods, namely logistic regression, on datasets to assess and investigate whether there are any indicators or chances of cancer, and if the individual is found to be malignant, they also estimate the stage of cancer[2].

Radhika P R et al. (2018): Lung cancer is defined as the development of malignant cells in the lungs. Because of the increasing frequency of cancer, both men and women's death rates have increased. Lung cancer cannot be avoided, but it can be decreased. As a result, early identification of lung cancer is critical for patient survival. The number of persons afflicted by lung cancer is directly proportionate to the number of chain smokers. Classification algorithms such as Naive Bayes, SVM, Decision Tree, and Logistic Regression were used to predict lung cancer. The primary goal of this research is to investigate the performance of classification algorithms in the early detection of lung cancer[3].

Prabadevi et al. (2020): Cancerous breast cells are the most promising of all malignancies that are prevalent among women and the leading cause of mortality

worldwide. Accurate detection of these cancer cells in their early stages is critical, which may be accomplished using various data mining and machine learning approaches. As a result, a comparison of several machine learning approaches is carried out. The WEKA tool is used to determine it. Furthermore, the chosen machine learning methods are assessed based on prediction accuracy, and performance comparisons of each classifier using a ROC curve on several classifiers are done[4].

Jabbar et al. (2020): "Lung Cancer" is claimed to be prominent cause of cancer-related death worldwide. As a result, early detection, prediction, and diagnosis of lung cancer have become critical since it expedites and streamlines the subsequent clinical board. Because of their accuracy, machine learning techniques have been used to accelerate the progression and treatment of malignant illnesses. Various forms of machine learning algorithms have been used in the healthcare industry for lung cancer analysis and prognosis. The elements that cause lung cancer and the use of Machine Learning methods are examined in this review article, with specific emphasis on their respective strengths and drawbacks[5].

Hazra et al. (2017): Lung cancer is one of the most prevalent and major causes of cancer mortality in humans. The advanced detection of cancer is the primary factor in increasing a patient's chances of survival. This research examines and compares the performance of the support vector machine (SVM) and logistic regression (LR) algorithms in predicting the survival rate of lung cancer patients. These approaches have been used to detect the chances of survival for lung cancer patients and to assist clinicians in making judgements about the disease's prognosis[6].

Ravichandran et al.: The post-surgery survival rate for lung cancer is quite low, regardless of whether it is small cell or non-small cell. Many studies have been conducted using machine learning to estimate life expectancy after thoracic surgery for individuals with lung cancer. Based on UCI data sets, many machine learning models have been used to estimate post-thoracic surgery life expectancy. In addition, work has been done on attribute ranking and selection in order to improve prediction accuracy with machine learning algorithms. As a result, the authors devised a deep neural network-based technique in predicting post-thoracic life expectancy, which is the most

advanced kind of neural networks. This is based on a dataset produced using machine learning at the Wroclaw Thoracic Surgery Centre[7].

Jurka et al. (2018): Lung cancer is a prevalent cancer with a poor five-year survival rate. Although the electronic nose has lately been examined as a possible ideal screening tool for early detection of lung cancer, no statistical technique has been proposed as the preferred one. The study's goal was to look at using logistic regression analysis to assess patients' exhaled breath samples using an electronic nose in order to distinguish lung cancer patients from patients with other lung disorders and healthy people. The study included 252 cancer patients and 223 non-cancer patients with histologically or cytologically confirmed untreated lung cancer, patients with other lung illnesses, and healthy volunteers[8].

Gültepe et al. (2020): Lung cancer claimed the lives of 1.76 million individuals globally in 2018. The majority of these deaths are the result of late diagnosis, and early-stage diagnosis considerably enhances the probability of effective lung cancer therapy. In this regard, they conducted a study on machine learning approaches to improve lung cancer classification accuracy using 32 x 56 sized numerical data from the University of California's Machine Learning Repository website. The precision of the classification model was raised in this study by the efficient use of pre-processing approaches rather than the direct use of classification algorithms. To accomplish this improvement, nine datasets were created using pre-processing approaches, and six machine-learning classification algorithms were applied. The study's findings indicate that the k-nearest neighbours technique is accurate [9].

3. Objectives of Study:

A beautiful use of machine learning can reduce the load of doctors in detection of diseases and its different objects. This study is carried out to —

a. Detect the existence of lung cancer in the lung cancer dataset.
b. Detect the gender of the liver cancer in the liver cancer dataset.
c. Compare the predicted cases with the observed cases.
d. Resemble machine learning with astrologer’s angles.

4. Methods and Material:

Cancer is becoming one of the most common diseases afflicting individuals worldwide. There are other forms of cancer, but we are dealing with lung and liver cancer here. Lung cancer is defined as the development of malignant cells in the lungs. Lung cancer is a condition that causes uncontrollable cell growth in the lungs. Machine learning is the broad term for computer algorithms that model a problem based on its data. It can enhance estimate of outcomes by using categorization algorithms based on structured data in predefined categories. Machine learning has had a tremendous influence on the biomedical area for the prediction and detection of liver disease. Machine learning can quickly answer medical problems and lower diagnostic costs.

- **Materials**

This study relies on secondary data. For this study, data from lung cancer patients and Indian liver patients were collected. These datasets were obtained from http://www.kaggle.com.

Lung cancer data:

This dataset contains the following information--

The total number of qualities is 16

Number of occurrences: 284

Information about attributes:

M (male), F (female) (female), Age: The patient's age, YES = 2, NO = 1 for smoking YES = 2, NO = 1, yellow fingers Anxiety: YES = 2, NO = 1. Peer pressure: YES = 2 and NO = 1. Chronic Illness: YES = 2, NO = 1. YES = 2, NO = 1 for fatigue YES = 2, NO = 1 for allergy. YES = 2, NO = 1 for wheezing YES = 2, NO = 1 for alcohol. YES = 2, NO = 1 for coughing Breathing Shortage: YES = 2, NO = 1. Difficulty Swallowing: YES = 2, NO = 1. YES = 2, NO = 1 for chest discomfort. Yes or no to lung cancer.

Table 1: Lung cancer dataset

ID	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHESTPAIN	LUNG_CANCER
0	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
1	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
2	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
3	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
4	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
30	F	56	1	1	1	2	2	2	1	1	2	2	2	2	1	YES

ID	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHESTPAIN	LUNG_CANCER
4																
305	M	70	2	1	1	1	1	2	2	2	2	2	2	1	2	YES
306	M	58	2	1	1	1	1	1	2	2	2	2	1	1	2	YES
307	M	67	2	1	2	1	1	2	2	1	2	2	2	1	2	YES
308	M	62	1	1	1	2	1	2	2	2	2	1	1	2	1	YES

Source: https://www.kaggle.com

- **Indian liver patients data set**

This data collection covers liver patient records as well as non-liver patient records obtained in Andhra Pradesh's North East. When Liver Problem = 1, the individual is a liver patient.

Liver Problem = 2 indicates that the person is not a liver patient.

Data Description: Patient's age, Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins Albumin, Albumin/Globulin Ratio.

Dataset: Indian liver patients

Table 2: Indian liver patient dataset

ID	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin	Albumin_and_Globulin_Ratio	Dataset
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.00	1

4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.40	1
...	...	...	...	...	...	...	...	...	...	...	...
574	60	Male	0.5	0.1	500	20	34	5.9	1.6	0.37	2
575	40	Male	0.6	0.1	98	35	31	6.0	3.2	1.10	1
576	52	Male	0.8	0.2	245	48	49	6.4	3.2	1.00	1
577	31	Male	1.3	0.5	184	29	32	6.8	3.4	1.00	1
578	38	Male	1.0	0.3	216	21	24	7.3	4.4	1.50	2

Source: https://www.kaggle.com

- **Methods**

The success of cancer prediction systems enables people to learn about their cancer danger at a cheap cost and to make appropriate decisions based on their cancer threat status. Cases of liver disease have been steadily increasing as a result of excessive alcohol intake, consumption of harmful feasts, and consumption of contaminated food, pickles, and

medications. In order to lessen the strain on doctors, a data scientist can create a predictive machine literacy that can predict whether or not a person has a liver disease. The goal for a data scientist is to create a logistic machine literacy model that predicts if a patient is healthy (non-liver case) or unwell (liver case) based on clinical and epidemiological data. The goal for a data scientist is to create a logistic machine literacy model that predicts if a case is healthy (non-liver case) or unwell (liver case) based on clinical and demographic data. The training and testing are carried out in a Python environment utilising the Jupyter platform. The codes used for this purpose are given in Appendix-I:

5. Results and Discussion

In this section, the results of the materials mentioned in the chapter II are shown and discussed accordingly.

- **Results**

In case of the lung cancer dataset, the results obtained are shown in table 3:

Table 3: Predicted results vs observed results

ID	Observed	Predicted	Remarks
83	Yes	Yes	Correct
224	Yes	Yes	Correct
269	Yes	Yes	Correct
245	Yes	No	Wrong
⋮	⋮	⋮	⋮
145	Yes	Yes	Correct
174	Yes	Yes	Correct
126	Yes	Yes	Correct
6	No	Yes	Wrong
277	Yes	Yes	Correct
14	No	Yes	Wrong

The accuracy of the result of lung cancer dataset is obtained to be 92%. From table 3 als o it can be understood. In case of the Indian liver patient's dataset, the results so obtaine d are depicted in table:

Table 4: Predicted results vs observed results

ID	Observed	Predicted	Remarks
341	Male	Male	Correct
121	Male	Male	Correct
575	Male	Male	Correct
335	Male	Male	Correct
54	Male		
⋮	⋮	⋮	⋮
180	Male	Male	Correct
270	Male	Male	Correct
159	Male	Male	Correct
319	Female	Male	Wrong
468	Male	Male	Correct

The accuracy of the result of lung cancer dataset is obtained to be 75.11%. From table 4 also it can be understood.

6. Discussion

In table 3, the predicted results are obtained, depicted and compared with the observed cases of lung cancer list. In this table, the test cases are predicted. It has given 92% accuracy. In table 4, the predictions for the Indian liver cases are penned down. These results are also compared with the observed cases. The accuracy in this case has been obtained to be 75.11%.

7. Conclusion

In this section, the different points derived out from the entire study are highlighted. Several methods are available to study this phenomenon. Our contribution towards detecting the categories are to resembling machine learning with astrologer. Astrologers predict the future of a person or the category of the object based on some information. Here also we have given some of the information to the machine to train itself and based on these training data, the machine predicts the test data. This concept is like the concept of an astrologer. In this particular study, we have used ‘train_test_split’ algorithm to train the machine taking a test size of 20%. This has been followed by logistic regression model to fit the machine for prediction. In our study, in case of lung cancer, the predictions are found to have 92% accuracy rate. In case of liver patient, 75.11% accuracy rate is depicted.

- **Future Scope:** In future, the study will be carried forward to clinical and image data. This can reduce the load of the doctors.

- **Appendix-I**

```
# lung cancer dataset
import pandas as pn
lung= pn.read_csv('survey lung cancer_1.csv')
lung
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(lung[['AGE','SMOKING','ANXIETY','PEE
R_PRESSURE','CHRONICDISEASE','ALCOHOL
CONSUMING','COUGHING']],lung.LUNG_CANCER,test_size=0.2)
from sklearn.linear_model import LogisticRegression
log = LogisticRegression()
log.fit(x_train,y_train)
log.score(x_train,y_train)
# Indian liver patient's dataset
import pandas as pn
ind_liv=pn.read_csv('indian_liver_patient.csv')
ind_liv.head()
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test =
train_test_split(ind_liv[['Age','Total_Bilirubin','Direct_Bilirubin','Alkaline_Phosphotase','
Alamine_Aminotransferase','Aspartate_Aminotransferase','Total_Protiens','Albumin','Alb
umin_and_Globulin_Ratio','Dataset']],ind_liv.Gender)
from sklearn.linear_model import LogisticRegression
loglv = LogisticRegression()
loglv.fit(x_train,y_train)
loglv.predict(x_test)
```

References

[1] Y. Geltepe, "Performance of Lung Cancer Prediction Methods Using Different Classification Algorithms," *Computers, Materials & Continua*, vol. 67, no. 2, pp. 2015–2028, 2021, doi: 10.32604/cmc.2021.014631.

[2] Student of SRMIST, Ramapuram, Chennai, India. *et al.*, "Detection Analysis of Various Types of Cancer by Logistic Regression using Machine Learning," *IJEAT*, vol. 9, no. 1, pp. 99–104, Oct. 2019, doi: 10.35940/ijeat.A1055.109119.

[3] R. P.R., R. A. S. Nair, and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, India, Feb. 2019, pp. 1–4. doi: 10.1109/ICECCT.2019.8869001.

[4] B. Prabadevi, N. Deepa, L. B. Krithika, and V. Vinod, "Analysis of Machine Learning Algorithms on Cancer Dataset," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Vellore, India, Feb. 2020, pp. 1–10. doi: 10.1109/ic-ETITE47903.2020.36.

[5] S. S. Raoof, M. A. Jabbar, and S. A. Fathima, "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Bangalore, India, Mar. 2020, pp. 108–115. doi: 10.1109/ICIMIA48430.2020.9074947.

[6] A. Hazra, N. Bera, and A. Mandal, "Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms," *IJCA*, vol. 174, no. 2, pp. 19–24, Sep. 2017, doi: 10.5120/ijca2017915325.

[7] K. Gulati, S. Saravana Kumar, R. Sarath Kumar Boddu, K. Sarvakar, D. Kumar Sharma, and M. Z. M. Nomani, "Comparative analysis of machine learning-based classification models using sentiment classification of tweets related to COVID-19 pandemic," *Materials Today: Proceedings*, vol. 51, pp. 38–41, Jan. 2022, doi: 10.1016/j.matpr.2021.04.364.

[8] M. Tirzïte, M. Bukovskis, G. Strazda, N. Jurka, and I. Taivans, "Detection of lung cancer with electronic nose and logistic regression analysis," *J. Breath Res.*, vol. 13, no. 1, p. 016006, Nov. 2018, doi: 10.1088/1752-7163/aae1b8.

[9] S. R. Pegu, M. Choudhury, S. Rajkhowa, D. Kumar, and B. R. Gulati, "Molecular characterization and pathological studies of Japanese encephalitis virus in pigs of Assam," *Journal of Entomology and Zoology Studies*, p. 5.

