ISSN NO. 2091-1742

SCITECH

Journal of Scientific and Technical Studies
Nepal Engineering College, Pokhara University

Research Article

DOI:10.3126/scitech.v17i1.60492

Machine-Learning Based Prediction of Lung Cancer

Trailokya Raj Ojha[1*], Menuka Maharjan[2]
Department of Computer Science and Engineering
Nepal Engineering College, Nepal
Email: trailokyaro@nec.edu.np[1], menukam@nec.edu.np[2]
*Corresponding Author

Received: April 20, 2023
Accepted: October 10, 2023

Abstract

The incidence of lung cancer has now exceeded all other types of cancer globally, making it the leading cause of cancer-related deaths. Compared to other types of cancer, lung cancer has a poor prognosis and high mortality rate, making it challenging for humans to accurately predict its incidence rates. The goal of the study is to implement machine learning techniques for the early detection of lung cancer, which can increase patient survival rates. To find the most important factors and predict the possibility of lung cancer, a variety of data mining approaches, including logistic regression, k-means, and apriori algorithms are used in this study. Age, gender, smoking habit, and medical history are a few of the factors included in the dataset used for the study. The logistic regression classifier shows an accuracy of 95% for the classification of lung cancer in patients. The simulation results obtained from the k-means clustering algorithm shows that the main causes for the possible occurrence of lung cancer are chronic diseases, fatigue, allergy, wheezing, alcohol consumption habit, and breath problem. Similarly, according to the association rule's findings, there is no chance of lung cancer developing in a non-drinker who is free of peer pressure, allergies, and wheezing issues.

Keywords: Apriori; classification; data mining; k-means; logistic regression; lung cancer; machine learning; prediction

Introduction

Lung cancer is considered a leading cause of cancer-related deaths globally; with many people succumbing to the disease every year. This is largely due to increased smoking habits and air pollution; which have made lung cancer a serious and ongoing global issue. People who have had previous lung diseases are at a higher risk of developing lung cancer. Smoking and tobacco consumption are considered the primary factors responsible for lung cancer. Detecting lung cancer early can greatly improve the chances of successful treatment and save lives.

As one of the major causes of death worldwide; cancer begins when a few cells in a body start to grow in an uncontrolled manner. Every type of tumor arises through "hereditary adjustments" and "epigenetic changes" to the DNA genome; even though various tumor types differ in how their cells proliferate and spread. Recent studies have strengthened the argument that epigenetic changes are important to the emergence of human malignancies (Kumar & Rao; 2021). According to Singh & Gupta (2019); an estimated 1.7 million people die due to lung cancer. In approximately 80% of lung cancer cases worldwide; smoking is considered a primary cause. Including smoking; many other factors are also responsible for causing lung cancer.

Different types of machine learning algorithms can predict the relationship between risk factors and lung cancer. They provide precise analysis and predictions. In the last twenty years; artificial intelligence and machine learning algorithms have become increasingly essential in analyzing complex data and reaching reliable conclusions. Machine learning algorithms have been created to classify; forecast; or reduce the amount of raw data. In this study; we have used different machine learning algorithms to predict the possibility of the occurrence of lung cancer based on the participant's medical history and statistical data. The different machine learning algorithms used in this study are Logistic Regression; K-Means; and Apriori algorithms for classification; clustering; and association respectively.

The Kaggle repository was used to collect the dataset for this study (Kaggle; 2022). The dataset was pre-processed so that machine learning algorithms could use it. Using the data balancing technique; the unequal distribution of data across cancer and non-cancer classes was addressed. To determine the most accurate prediction model; the accuracy of each machine learning model is calculated and compared.

Related Works

Lung cancer is the major cause of cancer-related deaths globally; which is a serious and common disease. For patients to receive better care and experience better results; early detection and correct prognosis of lung cancer are crucial. Machine learning techniques have been used in recent years to categorize and predict lung cancer. Many machine-learning approaches are based on neural networks. Using labels from the original training data; this method entails categorizing the data into different groups. Several machine learning techniques; which may be roughly categorized as supervised or unsupervised methods; can be used to categorize data.

A unique multi-layered approach is proposed to construct a cancer risk prediction system that combines decision tree and clustering approaches by Ramachandran et al. (2014). The authors built a model that incorporates a decision tree as a classification algorithm and k-means as a clustering algorithm to detect cancer. A study conducted by Abdullah et al. (2021) examines the precision ratios of three classifiers; including support vector machine (SVM); k-nearest neighbor (k-NN); and convolutional neural network (CNN); in data taken from the UCI repository to find lung cancer at an early stage and possibly save many lives. The results of the experiment indicate that the SVM; which has an accuracy of 95.56%; obtains the best result; followed by CNN (92.11%) and KNN (88.40%).

In a study by Zhou et al. (2021); lung nodules on computed tomography (CT) images were accurately classified using deep learning approaches. High sensitivity and specificity were attained for their classification model; demonstrating the promise of machine learning techniques to assist in the diagnosis of lung cancer.

A study conducted by Tuncal et al. (2020) used support vector regression; a back-propagation learning algorithm; and a long-term memory network to analyze lung cancer data for eleven European countries with records going back to 1970. The findings show that all algorithms are capable of scoring highly when estimating incidence rates; although Support Vector Regression did better than the other approaches that were taken into account.

Danjuma (2015) conducted a study to identify and evaluate the performance of machine-learning algorithms in patients with lung cancer. On datasets for thoracic surgery adopted from the UCI repository; multilayer perceptron; J48; and the Naïve Bayes algorithms were utilized to train and evaluate models. To evaluate the performance of the classifier; a 10-fold cross-validation technique was used. With an accuracy of 82.3%; the comparative analysis reveals that the multilayer perceptron shows the best performance.

Kourou et al. (2015) examined several machine learning algorithms to assess the efficacy of ML techniques in cancer prognosis and prediction. They concluded that supervised models should be the primary subject of research while developing prediction algorithms. While Malvezzi et al. (2014) used a linear regression model to anticipate cancer mortality rates across the European Union; Ribes et al. (2014) employed Bayesian models to forecast both incidence and mortality rates in Catalonia. Random forest and Rule Induction Algorithms were used by Alhaj and Maghari (2017) to identify the cancer survival rates in the Gaza Strip.

Radhika et al. (2019) conducted a study to predict and categorize medical imaging data as a primary area of study. Comparative studies utilizing various machine learning methods revealed that the support vector

machines had greater accuracy (99.2%). Naive Bayes offers 87.87% accuracy; Decision Tree offers 90%; and Logistic Regression offers 66.7%. The data set for the study was adopted from the UCI repository.

Another study by Sujitha and Seenivasagam (2021) proposed a machine learning-based model for predicting the survival rate of lung cancer patients. They used data from the Surveillance; Epidemiology; and End Results (SEER) program and found that their model outperformed traditional statistical models in predicting survival rates.

The categorization and prediction of lung cancer have shown considerable potential for data mining approaches. These methods can increase the precision and efficiency of diagnosis; which will benefit patients. However; more investigation is required to confirm the reliability and accuracy of these models as well as their application in clinical practice.

Materials and Methods

This section describes the methodology used in this study and it is divided into three sub-sections namely data description and preprocessing; algorithm description; and implementation procedure.

Data description and pre-processing

The dataset used for the study purpose is adopted from the Kaggle repository. There were 309 entries in total; including 162 male and 147 female. One target class and 15 attributes make up the data set. The attributes of the data set are listed in Table 1.

Table 1: Data set description

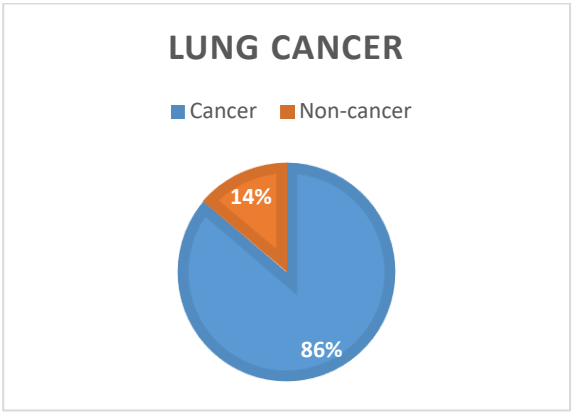
Attribute Name	Variable Type	Possible Value
Gender	Predictor Variable	Male; Female
Age	Predictor Variable	21 to 87
Smoking	Predictor Variable	Yes; No
Yellow_finger	Predictor Variable	Yes; No
Anxiety	Predictor Variable	Yes; No
Peer_pressure	Predictor Variable	Yes; No
Chronic diseases	Predictor Variable	Yes; No
Fatigue	Predictor Variable	Yes; No
Allergy	Predictor Variable	Yes; No
Wheezing	Predictor Variable	Yes; No
Alcohol consuming	Predictor Variable	Yes; No
Coughing	Predictor Variable	Yes; No
Shortness of breath	Predictor Variable	Yes; No
Swallowing difficulty	Predictor Variable	Yes; No
Chest pain	Predictor Variable	Yes; No
Lung_cancer	Response Variable	YES; NO

The data set available in the data repository might not be balanced or complete. The missing values and noise in data may affect negatively in final prediction. The data preprocessing technique is implemented to balance the dataset to achieve better accuracy. At this point; the null values are verified; and corrected. The preprocessing processes for the dataset include feature selection; value reduction; and normalization.

The first step in data preprocessing is checking for null values and if occurs it was filled by using the supervised filter in WEKA. In the next step; the numeric values were converted to categorical values as most of the values are in numeric form in the original data set. In original data set ‘2’ was used to represent ‘YES’ and ‘1’ was used for ‘NO’. For this; we have replaced the value '2' with 'Yes' and '1' with 'No’. To make it more understandable; the values of age attribute 'M' was replaced by 'male'; and 'F' was replaced with 'female'.

The target variable (lung cancer) column in the study's dataset had a value of "YES" in 270 rows whereas "NO" was present in just 39 rows; indicating a substantial amount of imbalance. If such uneven data is not regulated; predictions and outcomes are useless. 33 duplicate entries in the data set were also deleted. The dataset now

contains 276 entries; 238 entries with cancer; and 38 entries without cancer after the duplicate items have been removed. To balance the uneven distribution of cancer and non-cancer classes in the data set; SMOTE technique was applied in this study to oversample the minority class "non-cancer". The distribution of the data set based on the response variable before and after sampling is shown in Figure 1.



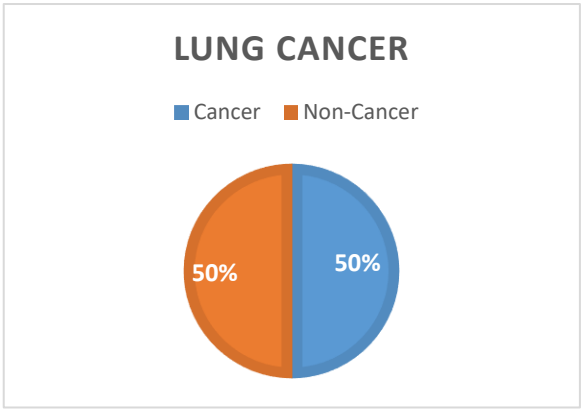


Figure 1 (a): Data before sampling

Figure 1 (b): Data after sampling

After balancing the data set; the next step is to develop the model. A test set and training set is created from the preprocessed data. A 10-fold cross-validation approach is used to find the performance of the machine learning algorithm. 90% of the total dataset is randomly chosen as the training set for the 10-fold cross-validation approach; while the remaining 10% is chosen as the test set. After dividing the dataset; we utilized classification; clustering; and association methods to train the model.

Algorithm description

Logistic Regression

Regression methodology known as logistic regression foresees a categorical dependent variable. To assess the statistical significance of the variables; the logistic regression equation includes the maximum likelihood ratio (Hosmer et al. 2013). The main objective of logistic regression is to estimate the probability of the response variable based on the values of one or more predictor variables or features.

Based on the values of a collection of predictor variables; logistic regression is effective at predicting the existence or absence of a characteristic or outcome. It is useful for models when the dependent variable is categorical even though it resembles a linear regression model (Kurt et al. 2008). Since there are only two possible values for the output attribute in the data set; logistic regression was selected in this instance.

The probability of the response variable is produced by the logistic regression model; which is a linear model that has been converted using a logistic or sigmoid function. Any real-valued input is translated by the logistic function; which is an S-shaped curve; into a probability between 0 and 1 (Hosmer et al. 2013). A maximum likelihood strategy is used to estimate the logistic regression model; to identify the parameter values that maximize the likelihood of the observed data.

K-means Algorithm

In unsupervised machine learning; the k-means algorithm is a common clustering algorithm. A set of data points is divided into k clusters using k-means; where k is the predetermined number of clusters. The goal of clustering is to find commonalities and designs in large data sets by dividing them into groups (Mohamad & Usman 2013). The k-means algorithm can decrease the cost function of clustering through the utilization of a distance function that matches categorical items; modes (rather than cluster means); and an intensity mechanism that enhances the modes during clustering. It is defined as a set X = X1;...; Xn of objects of n numeric values; a natural number k n; and a measured distance is given; the k-means algorithm seeks to divide X into m nonempty disjoint clusters Y1;...; Ym with Yi Yj = and m i=1 Yi = X so that the sum of the squared difference between data items and the data cluster centers is reduced (San et al. 2004).

Apriori Algorithm

The a priori algorithm is a group of steps for figuring out the much earlier forms set in a specific database. The data mining technique repeats the join and prune procedures until the most common item set is generated. The main goal of the apriori algorithm is to establish an association rule between different entities. The relationship between two or more variables is explained by the association rule.

In the first iteration; set X directly comprises the first selection item set Y1. If X = x1; x2... xn; then Y1=x1; x2;...; xn. First; the candidate item set Yk for the K-th iteration emerges from the common items Lk-1 found in the previous iteration. Then; for each item in the set; assign a counter with a starting value of zero and scan the data file Z in the correct sequential order. The counter for those sets will rise as long as each activity is present in each item set. After all; relationships have been scanned; the support level could be determined based on the real value of |Z| and the negligible support level of a particular Yk of the set of common items. Continue until there are no new items (Yabing 2013).

Implementation procedure

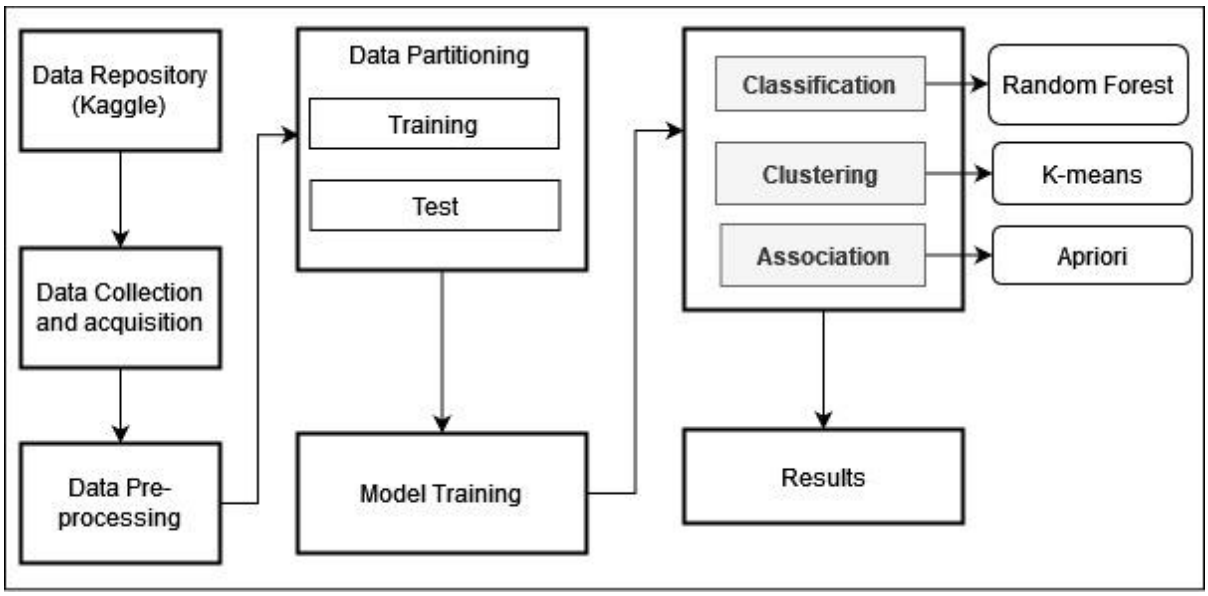


Figure 2: Working procedure

The proposed implementation procedure is illustrated in Figure 2. The implementation procedure involves different stages from data acquisition the result acquisition. The first step is data collection and acquisition following the preprocessing of data. Data partitioning is the next step. In data partitioning; we split data into training and test sets for model training. In model training; we have used classification; clustering; and association algorithms. For classification; the Logistic Regression algorithm is used; the k-means algorithm is used for clustering; and for association; apriori algorithm is used. After training the given model; we achieve the corresponding results; and the achieved result is represented according to the objective of the study.

Result and Discussion

The dataset adopted from the Kaggle repository was trained using WEKA 3.8.6 environment (WEKA 2022). A 10-fold cross-validation approach is used for training and testing the data set. A 90% of the total dataset is randomly chosen as the training set for the 10-fold cross-validation approach; while the remaining 10% is chosen as the test set. The details result of the 10-fold cross-validation for Logistic Regression is depicted in Figure 5 to Figure 8 in the supplementary section.

Classification Results

The process of classification involves developing a model that can categorize objects and predict their missing features. The outcome of the classification process can be described using different metrics; but for this paper; we have focused on measures such as accuracy; precision; recall; and f-measure. The result obtained by implementing logistic regression classifier is illustrated in Table 2.

Table 2: Performance measure of Logistic Regression Classifier

Classifier	Accuracy	Precision	Recall	F-Measure
Logistic Regression	0.950	0.954	0.950	0.952

Table 3: Confusion Metrics for the Logistic Regression Classifier

	a	b
a	226	12
b	11	227

a = Yes (the participant has the possibility of occurring lung cancer)

b = No (the participant does not have the possibility of occurrence of lung cancer)

The result obtained after training the model shows that the logistic regression classifier performed well in the detection of lung cancer with an accuracy of 95%. Other measures such as precision; recall; and f-measure for the classifier are measured at 95.4%; 95%; and 95.2% respectively. The result indicates that 95% of data has been classified correctly for the prediction of lung cancer and only 5% of data has been classified incorrectly.

The confusion metrics shown in Table 3 indicate that 226 instances were correctly classified as possibility of occurring lung cancer and 12 instances were incorrectly classified as they do not have the possibility of occurrence of lung cancer. Similarly; 227 instances were correctly classified as they do not have the possibility of occurrence of lung cancer; and 11 instances were incorrectly classified as they have the possibility of occurrence of lung cancer. In this way; a total of 95% of instances were correctly classified and only 5% of instances are incorrectly classified.

A receiver operating characteristic (ROC) curve is used as the graphic depiction of the classifier's results. The ROC curve for the logistic regression classifier is shown in Figure 3.

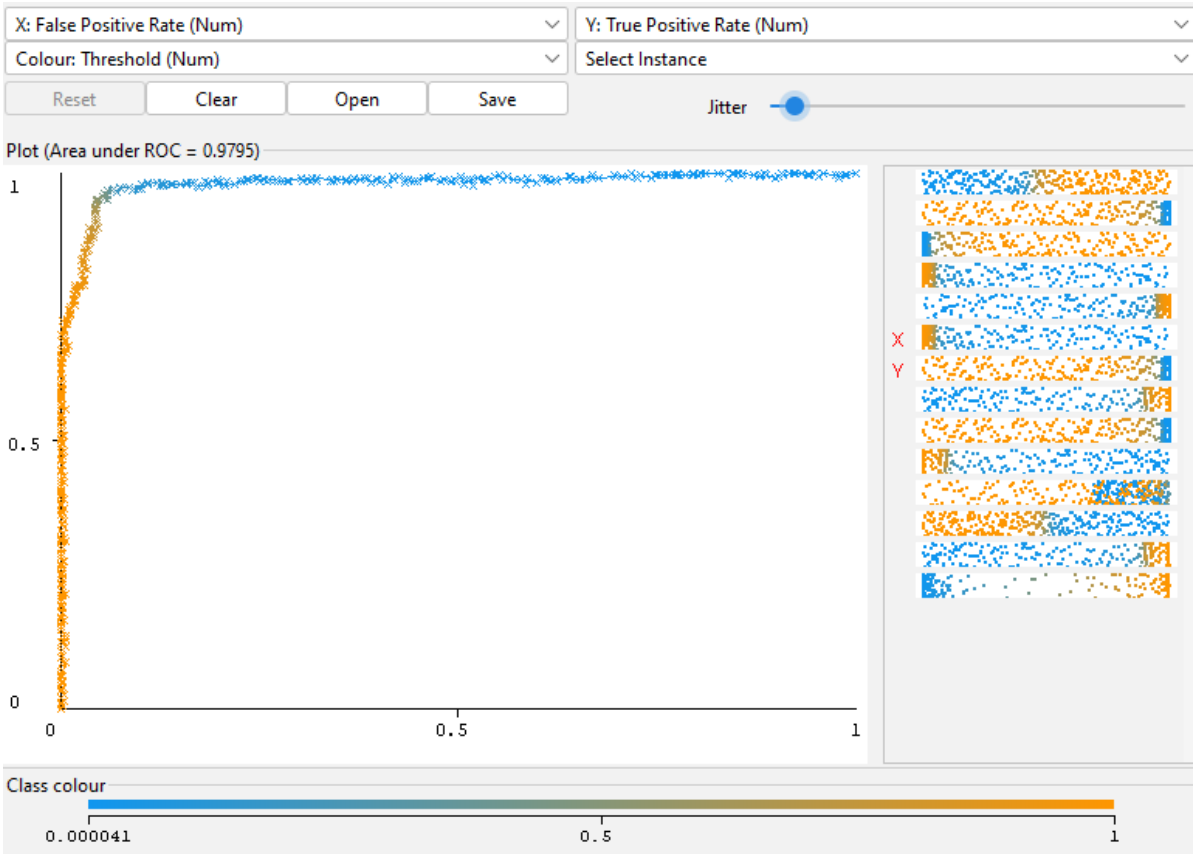


Figure 3: The ROC curve for the Logistic Regression classifier

The Receiver Operator Characteristic (ROC) curve is a measurement tool for binary classification problems. In essence; it distinguishes between "signal" and "noise" by comparing the true positive rate (TPR) against the false positive rate (FPR) at various threshold values. The area under the curve (AUC); used as a summary of the ROC curve; quantifies a classifier's ability to distinguish between classes. From the ROC curve for the logistic regression classifier; we can see that the value of AUC is 0.9795; this indicates that the classifier has performed in a better way to classify true positive and true negative cases.

Clustering Results

Clustering is the process of finding groups of things that are similar to one another yet distinct from other groups of objects (Prabha & Shanavas 2014). The k-means clustering algorithm was used in this study and the result obtained from the simulation is depicted in Table 4.

From the clustering result; we can conclude that the males with an average age of 64 having chronic diseases; fatigue; allergy; wheezing; alcohol consumption habit; and breath problem have the possibility of lung cancer. Similarly; females with an average age of 61 who does not have alcohol consumption habit; no chronic diseases; no peer pressure but have a smoking habit; yellow fingers; and fatigue does not have a chance of occurrence of lung cancer.

Table 4: Performance of k-means clustering algorithm

Attribute	**Full Data (476)**	**Cluster 0 (171)**	**Cluster 1 (305)**
Gender	Male	Male	Female
Age	62	64	61
Smoking	Yes	No	Yes
Yellow_fingers	Yes	No	Yes
Anxiety	No	No	No
Peer_pressure	No	No	No
Chronic diseases	No	Yes	No
Fatigue	Yes	Yes	Yes
Allergy	No	Yes	No
Wheezing	No	Yes	No
Alcohol Consuming	No	Yes	No
Coughing	No	Yes	No
Shortness of breath	Yes	Yes	Yes
Swallowing difficulties	No	No	No
Chest pain	No	Yes	No
Lung_cancer	***YES***	***YES***	***NO***

Association Results

The association data mining method can be applied to datasets to uncover intriguing correlations; recurrent patterns; interrelations; or causal structures between groups of elements. The association rule model was developed in WEKA using the apriori algorithm. The result obtained from the simulation is illustrated in Figure 4.

From the result obtained from the simulation; it is found that a non-alcoholic person without peer pressure and allergy; and not having a wheezing problem does not have any possibility of occurring lung cancer.

1. PEER_PRESSURE=No ALLERGY =No WHEEZING=No SWALLOWING DIFFICULTY=No 216 ==> ALCOHOL CONSUMING=No 216
2. PEER_PRESSURE=No ALLERGY =No SWALLOWING DIFFICULTY=No LUNG_CANCER=NO 218 ==> ALCOHOL CONSUMING=No 217
3. PEER_PRESSURE=No WHEEZING=No LUNG_CANCER=NO 217 ==> ALCOHOL CONSUMING=No 216
4. PEER_PRESSURE=No ALLERGY =No LUNG_CANCER=NO 223 ==> ALCOHOL CONSUMING=No 221
5. ALCOHOL CONSUMING=No LUNG_CANCER=NO 231 ==> ALLERGY =No 228

Figure 4: Performance of Apriori algorithm.

Conclusion

Data mining algorithms analyze a variety of variables; including patient medical history; and statistical data; to find patterns and connections that may not be immediately obvious to human observers. Different data mining techniques were used in this study to classify and predict the possibility of occurrence of the lung cancer in participants. The logistic regression classifier; k-means clustering algorithm; and apriori algorithm are used for the prediction of lung cancer. The Logistic Regression classifier has predicted the possibility of lung cancer in participants in a better way with an accuracy of 95%. The result obtained from the data mining techniques shows that males having chronic diseases; fatigue; allergies; wheezing; alcohol consumption habits; and breathing problems have a high chance of occurrence of lung cancer even if the participant does not have a smoking habit. Similarly; females are found to be safe from lung cancer. The main parameters that promote the occurrence of lung cancer are found to be peer pressure; allergy; alcohol consumption; and chronic diseases. This study helps medical personnel to predict and protect individuals from lung cancer.

This study was conducted using secondary data acquired from an online data repository. In the future; the study can be extended by using primary data collected directly from hospitals and health organizations and implementing the self-developed algorithms.

References

"Lung Cancer Dataset;" *https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection*. Accessed on 09 August 2022.

Abdullah; D.M.; Abdulazeez; A.M. and Sallow; A.B. 2021. Lung cancer prediction and classification based on correlation selection method using machine learning techniques. Qubahan Academic Journal 1(2); pp.141-149. https://doi.org/10.48161/qaj.v1n2a58.

Alhaj; M.A. and Maghari; A.Y. 2017; May. Cancer survivability prediction using random forest and rule induction algorithms. In *2017 8th International Conference on Information Technology (ICIT)* (pp. 388-391). IEEE.

Danjuma; K.J. 2015. Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients. *arXiv preprint arXiv:1504.04646*.

Hosmer Jr; D.W.; Lemeshow; S. and Sturdivant; R.X. 2013. Applied logistic regression (Vol. 398). John Wiley & Sons.

Kourou; K.; Exarchos; T.P.; Exarchos; K.P.; Karamouzis; M.V. and Fotiadis; D.I. 2015. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal 13; pp.8-17. https://dx.doi.org/10.1016/j.csbj.2014.11.005.

Kumar; M.S. and Rao; K.V. 2021; January. Prediction of Lung Cancer Using Machine Learning Technique: A Survey. In *2021 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-5). IEEE. https://dx.doi.org/10.1109/ICCCI50826.2021.9402320.

Kurt; I.; Ture; M. and Kurum; A.T. 2008. Comparing performances of logistic regression; classification; and regression tree; and neural networks for predicting coronary artery disease. Expert systems with applications 34(1); pp.366-374. http:/dx.doi.org/10.1016/j.eswa.2006.09.004.

Malvezzi; M.; Bertuccio; P.; Levi; F.; La Vecchia; C. and Negri; E. 2014. European cancer mortality predictions for the year 2014. Annals of Oncology 25(8); pp.1650-1656. https://dx.doi.org/10.1093/annonc/mdu138.

Mohamad; I.B. and Usman; D.; 2013. Standardization and its effects on K-means clustering algorithm. Research Journal of Applied Sciences; Engineering and Technology 6(17); pp.3299-3303. http://dx.doi.org/10.19026/rjaset.6.3638.

Prabha; S.L. and Shanavas; A.M. 2014. Educational data mining applications. Operations Research and Applications: An International Journal (ORAJ) 1(1); pp.23-29.

Radhika; P.R.; Nair; R.A. and Veena; G. 2019; February. A comparative study of lung cancer detection using machine learning algorithms. In *2019 IEEE International Conference on Electrical; Computer and Communication Technologies (ICECCT)* (pp. 1-4). IEEE.

Ramachandran; P.; Girija; N. and Bhuvaneswari; T. 2014. Early detection and prevention of cancer using data mining techniques. International Journal of Computer Applications 97(13). https://dx.doi.org/10.5120/17069-7492.

Ribes; J.; Esteban; L.; Cleries; R.; Galceran; J.; Marcos-Gragera; R.; Gispert; R. et al. 2014. Cancer incidence and mortality projections up to 2020 in Catalonia by means of Bayesian models. Clinical and Translational Oncology 16; pp.714-724. http://dx.doi.org/10.1007/s12094-013-1140-z.

San; O.M.; Huynh; V.N. and Nakamori; Y. 2004. An alternative extension of the k-means algorithm for clustering categorical data. International Journal of Applied Mathematics and computer science 14(2); pp.241-247.

Singh; G.A.P. and Gupta; P.K. 2019. Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. Neural Computing and Applications 31; pp.6863-6877. https://dx.doi.org/10.1007/s00521-018-3518-x.

Sujitha; R. and Seenivasagam; V.; 2021. Classification of lung cancer stages with machine learning over big data healthcare framework. Journal of Ambient Intelligence and Humanized Computing; 12; pp.5639-5649. http:/dx.doi.org/10.1007/s12652-020-02071-2.

Tuncal; K.; Sekeroglu; B. and Ozkan; C. 2020. Lung cancer incidence prediction using machine learning algorithms. Journal of Advances in Information Technology 11(2). https://dx.doi.org/10.12720/jait.11.2.91-96.

WEKA Tool. Available Online*:* *https://www.weka.io/*. Accessed on 27 March 2023.

Yabing; J. 2013. Research of an improved apriori algorithm in data mining association rules. International Journal of Computer and Communication Engineering 2(1); p.25.

Zhou; J.; Wang; W.; Lei; B.; Ge; W.; Huang; Y.; Zhang; L. et al. 2021. Automatic detection and classification of focal liver lesions based on deep convolutional neural networks: a preliminary study. Frontiers in oncology 10; p.581210. https://dx.doi.org/10.3389/fonc.2020.581210.

Supplementary Material

The details output of 10-fold cross-validation for the Logistic Regression classifier is shown in Figures 5 to 8. Basic run information for Logistic Regression is shown in Figure 5. Figure 6 depicts the training set description for 10-fold cross-validation for Logistic Regression. Odds ratios are shown in Figure 7 and the summary of the results is shown in Figure 8.

```
=== Run information ===

Scheme:       weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Relation:     lung cancer_processed data
Instances:    476
Attributes:   16
              GENDER
              AGE
              SMOKING
              YELLOW_FINGERS
              ANXIETY
              PEER_PRESSURE
              CHRONIC DISEASE
              FATIGUE
              ALLERGY
              WHEEZING
              ALCOHOL CONSUMING
              COUGHING
              SHORTNESS OF BREATH
              SWALLOWING DIFFICULTY
              CHEST PAIN
              LUNG_CANCER
Test mode:    10-fold cross-validation
```

Figure 5: Basic run information for Logistic Regression

```
=== Classifier model (full training set) ===

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                               Class
Variable                         YES
===================================
GENDER=Female                 1.2365
AGE                           0.0226
SMOKING=Yes                    2.383
YELLOW_FINGERS=No            -1.9578
ANXIETY=No                   -0.2635
PEER_PRESSURE=Yes             2.6797
CHRONIC DISEASE=Yes           4.4871
FATIGUE =No                  -3.3509
ALLERGY =Yes                  1.9347
WHEEZING=No                  -1.1332
ALCOHOL CONSUMING=No         -2.4093
COUGHING=No                  -3.9458
SHORTNESS OF BREATH=No        1.0187
SWALLOWING DIFFICULTY=No     -4.4183
CHEST PAIN=No                -0.7539
Intercept                     5.2815
```

Figure 6: Training set description for 10-fold cross-validation for Logistic Regression

```
Odds Ratios...
                                 Class
Variable                           YES
===================================
GENDER=Female                   3.4434
AGE                             1.0229
SMOKING=Yes                    10.8371
YELLOW_FINGERS=No               0.1412
ANXIETY=No                      0.7684
PEER_PRESSURE=Yes                14.58
CHRONIC DISEASE=Yes             88.8635
FATIGUE =No                     0.0351
ALLERGY =Yes                    6.9216
WHEEZING=No                      0.322
ALCOHOL CONSUMING=No            0.0899
COUGHING=No                     0.0193
SHORTNESS OF BREATH=No          2.7696
SWALLOWING DIFFICULTY=No        0.0121
CHEST PAIN=No                   0.4705
```

Figure 7: Odds Ratios

```
Time taken to build model: 0.1 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         453               95.1681 %
Incorrectly Classified Instances        23                4.8319 %
Kappa statistic                          0.9034
Mean absolute error                      0.0746
Root mean squared error                  0.2083
Relative absolute error                 14.9208 %
Root relative squared error             41.6558 %
Total Number of Instances              476

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.950     0.046     0.954      0.950    0.952      0.903    0.980     0.983     YES
               0.954     0.050     0.950      0.954    0.952      0.903    0.980     0.974     NO
Weighted Avg.  0.952     0.048     0.952      0.952    0.952      0.903    0.980     0.979

=== Confusion Matrix ===

   a   b   <-- classified as
 226  12 |   a = YES
  11 227 |   b = NO
```

Figure 8: Summary of the result for Logistic Regression