**Lung Cancer Prediction Through Machine Learning**

**Aadi Poddar Anjishnu Saw Dr. Mahendra Kumar Gourisaria**

*School of Computer Engineering School of Computer Engineering School of Computer Engineering*

*KIIT Deemed to be University KIIT Deemed to be University KIIT Deemed to be University*

*Bhubaneswar 751024, Odisha Bhubaneswar 751024, Odisha Bhubaneswar 751024, Odisha*

[*2305641@kiit.ac.in*](mailto:2305641@kiit.ac.in)[*23051895@kiit.ac.in*](mailto:23051895@kiit.ac.in)[*mkgourisaria2010@gmail.com*](mailto:mkgourisaria2010@gmail.com)

**Abstract.** Lung Cancer is one of the most common cancers in the world. Doctors and people are very cautious about predicting cancer based on their day-to-day habits so that it can be diagnosed as early as possible. This paper draws attention to the lung cancer predictions based on different objectives like gender and age of the person and their habits like smoking and alcohol consumption. The effectiveness of cancer prediction system helps the people to know their cancer risk with low cost and it also helps the people to take appropriate actions based on their cancer risk status. This paper involves various machine learning regression techniques like Linear Regression, Lasso Regression, Ridge Regression, Logistic Regression, Decision Tree and Random Forests. Here, all the above techniques are compared, and it is found that Logistic regression with all the features gives the best results.

**Keywords**- Lung Cancer, Linear Regression, Decision Tree Classifier, Logistic Regression, Random Forests

1. **Introduction**

Lung cancer is the leading cause of cancer death and the most diagnosed cancer. As per the data 1.8 million new cases were reported 2012, with 58% in less developed areas [1]. Most patients are diagnosed at an advanced stage, resulting in a low survival rate. In the past few years 14% of all caner diagnoses and 27% of all cancer deaths were due to Lung Cancer [2]. To solve this problem, machine learning engineers from across the world with data scientists are working to predict precise results to the doctors, patients and common people so that lung cancer can be predicted and diagnosed as early as possible. Predicting and determining cancer in an early stage is better and has a higher chance of cure. The patient has a higher rate of survival. Machine learning and generative modelling have been able to infiltrate various domains of scientific progress including economic forecasting and medicine. High dimensional modelling of big data has been detrimental to performing diagnostic and prognostic procedures. It is possible to understand attributes pertaining to organs of the human body vital for carrying out necessary life processes. The use of machine learning models significantly assists in prediction of lung cancer. The various aspects of Data Science and Machine Learning utilized were: Data Exploration, Machine Learning Model Developing, Machine Learning Model Engineering. Data is cleaned and sorted before training the models removing unnecessary data attributes which do not contribute in the classification and prediction process.

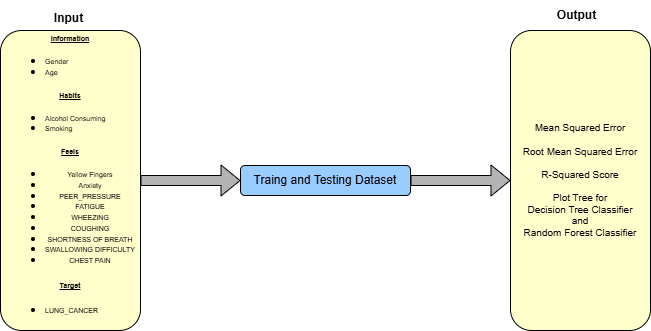
The models in this paper use variables regarding basic information about the person like gender and age. It consists of other attributes like habbits of the person like smoking and alcohol consumption, if the person is a regular smoker and drinker. It takes more information about how the person feels in day-to-day life like Fatigue, if the person feels exhausted. It takes variables regarding if the person feels wheezing in which they make a high-pitched sound while breathing. If the person has Shortness of Breath. If the person has Swallowing Difficulty, Chest pain, Yellow fingers and is under Peer Pressure. All these variables and attributes are combined to give the lung cancer result which is ther target variable.

For getting precise predictions, first, all measuring techniques like Negative Root Mean Square Error, Mean Squared Error and R2 Score are used on all the features for Linear Regression, Lasso Regression, Ridge Regression, Logistic Regression and Decision Tree.

1. **Data Preparation**

**The Dataset**

This section provides a brief description of the dataset used. The algorithms employed for predictions are based on real-world data, which is often difficult to handle, unbalanced, and incomplete. The dataset consists of 309 records and 16 attributes. Attributes include basic demographic information (Gender, Age), health-related symptoms (YELLOW\_FINGERS, ANXIETY, FATIGUE, WHEEZING, COUGHING, SHORTNESS OF BREATH, SWALLOWING DIFFICULTY, CHEST PAIN), habits (ALCOHOL CONSUMPTION, SMOKING), and additional factors (ALLERGY, PEER\_PRESSURE). The target attribute, LUNG\_CANCER, indicates the presence or absence of lung cancer.

****

**Preprocess Data**

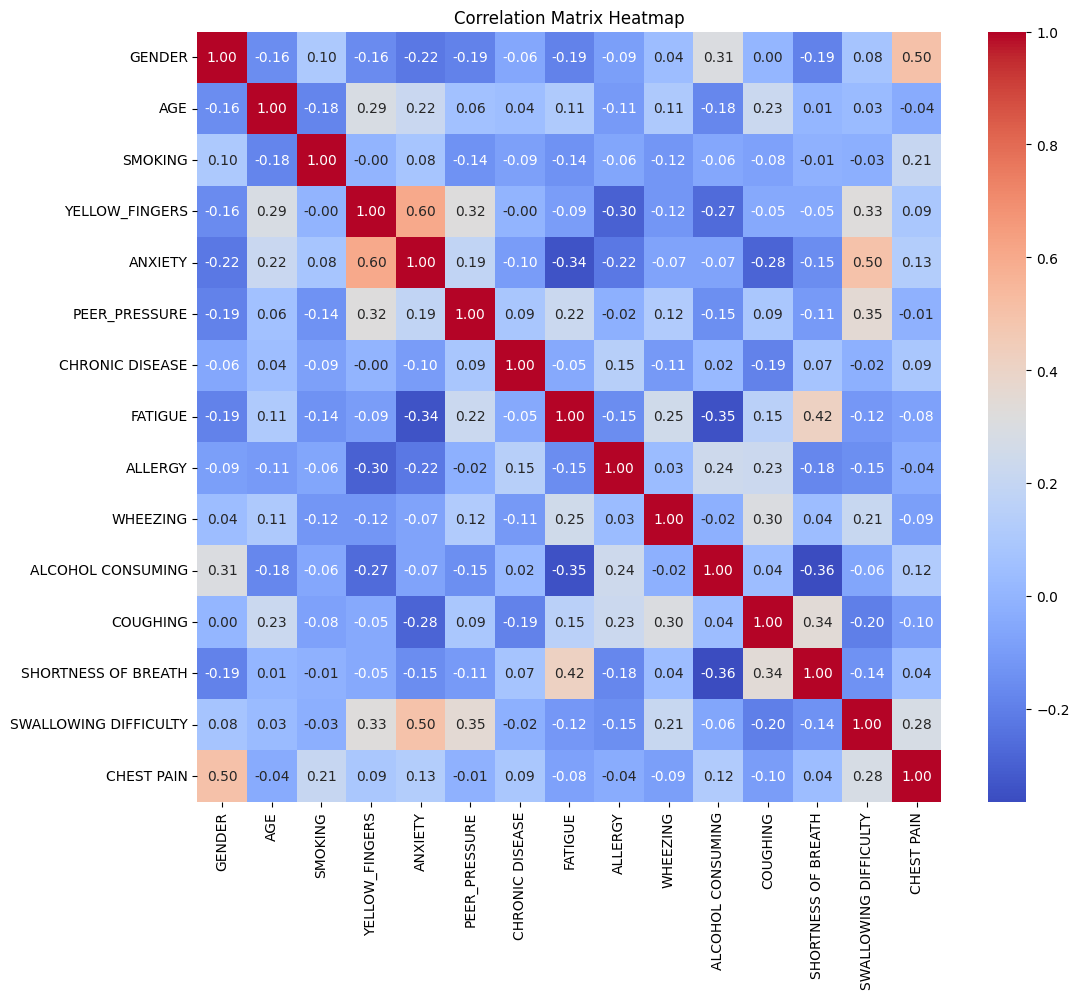
The Gender attribute, containing values M (male) and F (female), is replaced with numerical values. Categorical data is converted to binary by mapping M to 1 and F to 0. The target variable (LUNG\_CANCER) is encoded numerically, where 1 represents "YES" and 0 represents "NO." Binary encoding is ensured by replacing values such as 2 (originally denoting "YES" in features like smoking or anxiety) with 1, standardizing all features to a 0/1 binary format.

**Splitting the Data**

Input features (e.g., smoking, age, symptoms) are stored in one variable, while the target variable (LUNG\_CANCER) is stored separately. The dataset is split into training and testing sets, with 20% reserved for testing and 80% for training. The training set is used to train the machine learning model, and the testing set evaluates performance on unseen data. Dependent variables are predicted using the trained algorithm, and results are compared against original values to minimize error.

**Correlation Matrix**

The correlation matrix is used on the features. Highly correlated features with the target attribute are selected. They are taken into account to see the correlation between them. Highly correlated features are deselected [3][4].



1. **Technology**

**Linear Regression**

Linear Regression is used between the input attributes and the target variable to find a linear relationship between them [5][6][7]. We try to create a model with the training dataset which is a hypothesis that takes the new input attribute and gives the output of the target attribute. We use performance metrices to verify whether this model is performing well or not.

A model is created using the training dataset, forming a hypothesis that predicts the target variable from input features. Performance metrics are employed to validate the model. In Linear Regression, the best-fit line is determined to model the relationship between input features (x-axis) and the target variable (y-axis), expressed as:

***y = β₀ + β₁x₁ + β₂x₂ + ⋯ + βₙxₙ + ϵ***

***β₀*** is the Intercept. When x = 0 then y = ***β₀.*** This shows that at what point the line is meeting the y axis.

***β₁, β₂, …, βₙ*** are the slope or coefficients for the input features ***x₁, x₂, …, xₙ***

***ϵ*** denotes the error term that adds unwanted noise in the relationship between the regressors and the dependent variables.

Algorithm is best to use when we know that independent and dependent variables have linear relationship, but it is always assumed that there is a straight-line relationship dependent and independent attributes. Linear Regression is optimized for fast speeds.

**Overfitting**

* Model performs well with Training Data (Low Bias) but
* fails to perform well with Test Data (High Variance)

**Underfitting**

* Model Accuracy is bad with Training Data (High Bias)
* Model Accuracy is bad with Test Data (High Variance)

**Generalized Model**

* Model Accuracy is good with Training Data (Low Bias)
* Model Accuracy is good with Test Data (Low Variance)

**Ridge Regression (L2 Regularization)**

Ridge Regression is applied to prevent overfitting [8]. We do stop unless we get a line which be able to handle as a generalized model. It is used in such a way that we should never overfit. It introduces an L2 regularization term to the linear regression to ensure model stability.

(lambda) is the hyperparameter controlling the penalty strength.

are the coefficients.

Coefficients are shrunken towards zero. As we change the values we get the best fit line. After the line we get a slope which will decrease each time. After sometime this value will not decrease it will be a minimal and smaller value. We specify iterations that is how many times we have to train them. Iteration is also a hyperparameter. We cannot get to 0 because then it will be an overfitting model. It is best to use for datasets where overfitting is a risk. It always assumes that there is a straight-line relationship between. It Assumes Linear Relationships and is Ineffective for modeling non-linear interactions.

**Lasso Regression (L1 Regularization)**

Lasso Regression is similar to Ridge Regression. Lasso Regression adds the sum of mod slope instead of square of slope [9]. This Mod of slop work is to do feature selection. As we go on summing the values whichever features are not playing a significant role, the slope value will be extremely small. The entire feature is then neglected.

Ridge Regression does square of the slope which increases that particular value but because of the mod in Lasso Regression that value will not increase. In summary this neglects all features that are not at all important for this model. The major benefits of this model are to prevent overfitting and in cases where there are many features where many of them are not important. It performs feature selection.

***λ*** (lambda) is a hyperparameter which will be found out be cross validation.

The cost function is to be reduced in such a way that it will never become zero but will be reduced based on the **λ** and the slope value.

**Logistic Regression (Classification)**

Logistic Regression works best with binary data. Linear regression model is used to create a best fit linear line representing a linear relationship between the attributes [10]. Sigmoid activation function is used to squash the points. A mathematical formula is applied on top of this linear regression to squash the line. The cost function of logistic regression will always give a global minimum. Combining the graphs of two cost functions based on different values of y a gradient descend is achieved.

**Cost Function:**

**Convergence Algorithm:**

**L** is learning rate.

This function is repeated until convergence.

**Decision Tree Regression**

Decision Tree is a sort algorithm where data is split according to a certain parameter [11]. Feature are selected by Information gain. Decision Tree does not have a good time complexity for execution purpose. It splits all the features to its complete depth [12]. For training data, the accuracy is high and for test data the accuracy is low. It leads to low bias and high variance. This can be solved by pruning, but pruning is an extensive task for large datasets.

**Entropy**

is the probability of yes

is theprobability of no

Entropy will always come to 0 for a pure split. Features are taken and split happens based on their categories. For an impure split, further division of categories takes place. For small set of features Entropy is used.

**Information Gain**

is the Entropy of the root Node

is the Entropy of the Categories

Feature having the highest Information Gain will be chosen to start the splitting first.

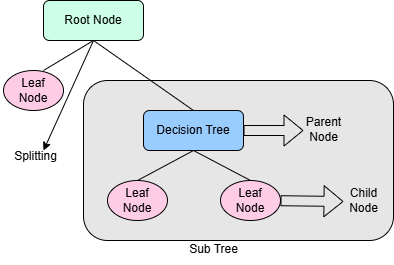
**Gini Impurity**

By default, the decision tree classifier uses Gini Impurity [13]

***n*** is the number of outputs (Yes, No)

***p*** is the probability of the output

Gini Impurity takes Less Time than Entropy.



**Random Forest Classifier and Regressor**

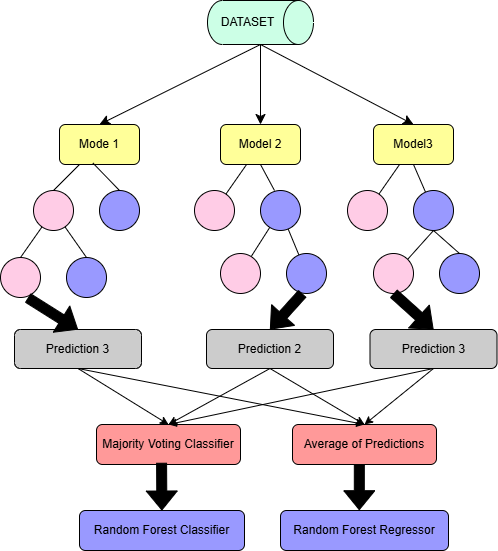
The main problem with Decision Tree is that it leads to overfitting when it is created without a hyperparameter [14][15]. The main aim of Random Forest is to convert the high variance of decision tree to low variance. Random Forest is a bagging technique in which a dataset has multiple models. Each model is a Decision Tree that is going to be used in Random Forest. Each Decision Tree model creates Low bias and High Variance. Combining each Decision Tree model in the form of bootstrap aggregator the High Variance will be converted to low variance. This is because majority of the voting will be taken from these decision trees. This is done by Majority Voting Classifiers [16].

In Random Forest each model is created by Row Sampling and Feature Sampling. This means that some set of rows and some sets of features are given to each model. This leads to features getting repeated, records and datapoints also get repeated in the models. Each model will be an expert in predicting the specific data on which it was trained.

On giving a test data to the Random Forest, each model gives its own specific output based on what it predicts. Majority Voting Classifiers technique is used to give the best prediction of the test data which combines the prediction of each model.

In Random Forest Regression the mean or average of the prediction of each model is taken instead of majority voting.

Normalization is not required in Random Forest and it is not impacted by Outliers.



1. **Evaluation Metrics**

Machine Learning is based on the analysis and testing of dataset, which on processing with algorithms, gives certain predictions. Accuracy is calculated which is how well the model predicted the results [17]. There will then be a comparison between the predicted results and the actual values to test on how accurate the model behaves. Error represents any false predictions made by the model [18]. For a regression model the evaluation metrics that are used are –

**Mean Squared Error (MSE)**

Mean Squared Error is the difference between the mean squared of the predicted values and the actual values [18]. It is the mean of the squared errors.

is the number of predictions

***x*** is the actual values

are the predicted values

**Root Mean Square Error (RMSE)**

Root Mean Square Error is the most used machine learning evaluation metrices. It is the square root of the Mean Squared Error. It is the square root of the mean of the squared errors [17]. The difference between the actual and predicted value is the error.

**Mean Squared Log Error (MSLE)**

Mean Squared Log Error is the ratio of predicted and actual value of the attributes.

**Score**

R square is the coefficient of determination. It varies from 0 to 1. It is the closeness of predicted data with actual data. It is represented by a percentage. Higher value of R square represents higher accuracy of the prediction model.

is the mean value of x

1. **Evaluation Metrics**

**Linear Regression**

Mean Square Error (MSE):

[0.06240464, 0.10083204, 0.08282308, 0.11788526, 0.07190797]

Mean of MSE:0.08717059843379187

R2 Score: 0.27453327218806

**Ridge Regression**

Mean Square Error (MSE):

[0.0639938, 0.09233465, 0.0805934, 0.12310473, 0.0650326]

Mean of MSE:0.08501183781577708

R2 Score: 0.40995768014421974

**Lasso Regression**

Mean Square Error (MSE):

[0.06291676, 0.09955315, 0.08700388, 0.11905419, 0.0705677]

Mean of MSE:0.08781913517914805

R2 Score: 0.3091275136656959

**Logistic Regression**

Best Score:0.9508368740754729

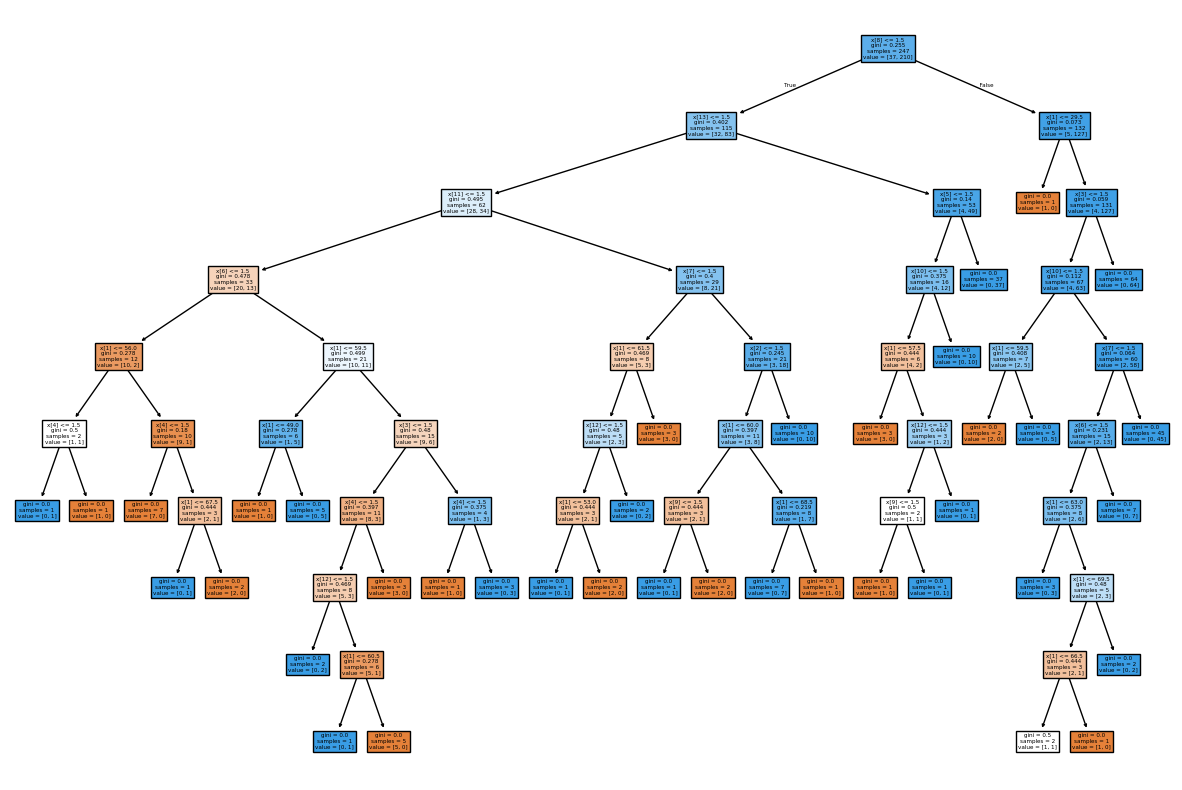
Accuracy: 98.38709677419355

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| No Cancer | 1.00 | 0.50 | 0.67 | 2 |
| Cancer | 0.98 | 1.00 | 0.99 | 60 |
| Accuracy |  |  | 0.98 | 62 |
| Macro Average | 0.99 | 0.75 | 0.83 | 62 |
| Weighted Average | 0.98 | 0.98 | 0.98 | 62 |

**Decision Tree Classifier**

Accuracy: 96.7741935483871

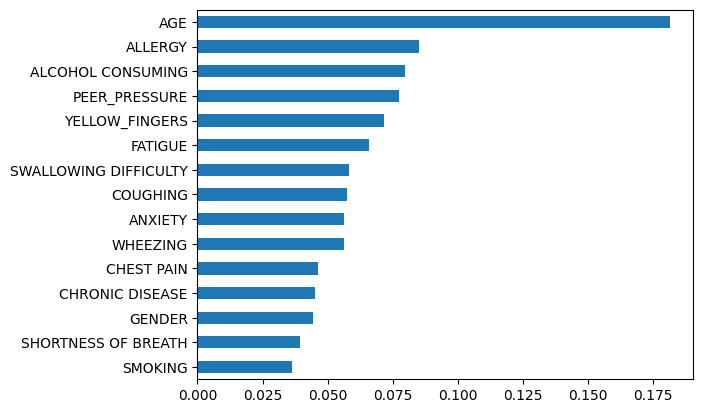
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| No Cancer | 0.50 | 0.50 | 0.50 | 2 |
| Cancer | 0.98 | 0.98 | 0.98 | 60 |
| Accuracy |  |  | 0.97 | 62 |
| Macro Average | 0.74 | 0.74 | 0.74 | 62 |
| Weighted Average | 0.97 | 0.97 | 0.97 | 62 |

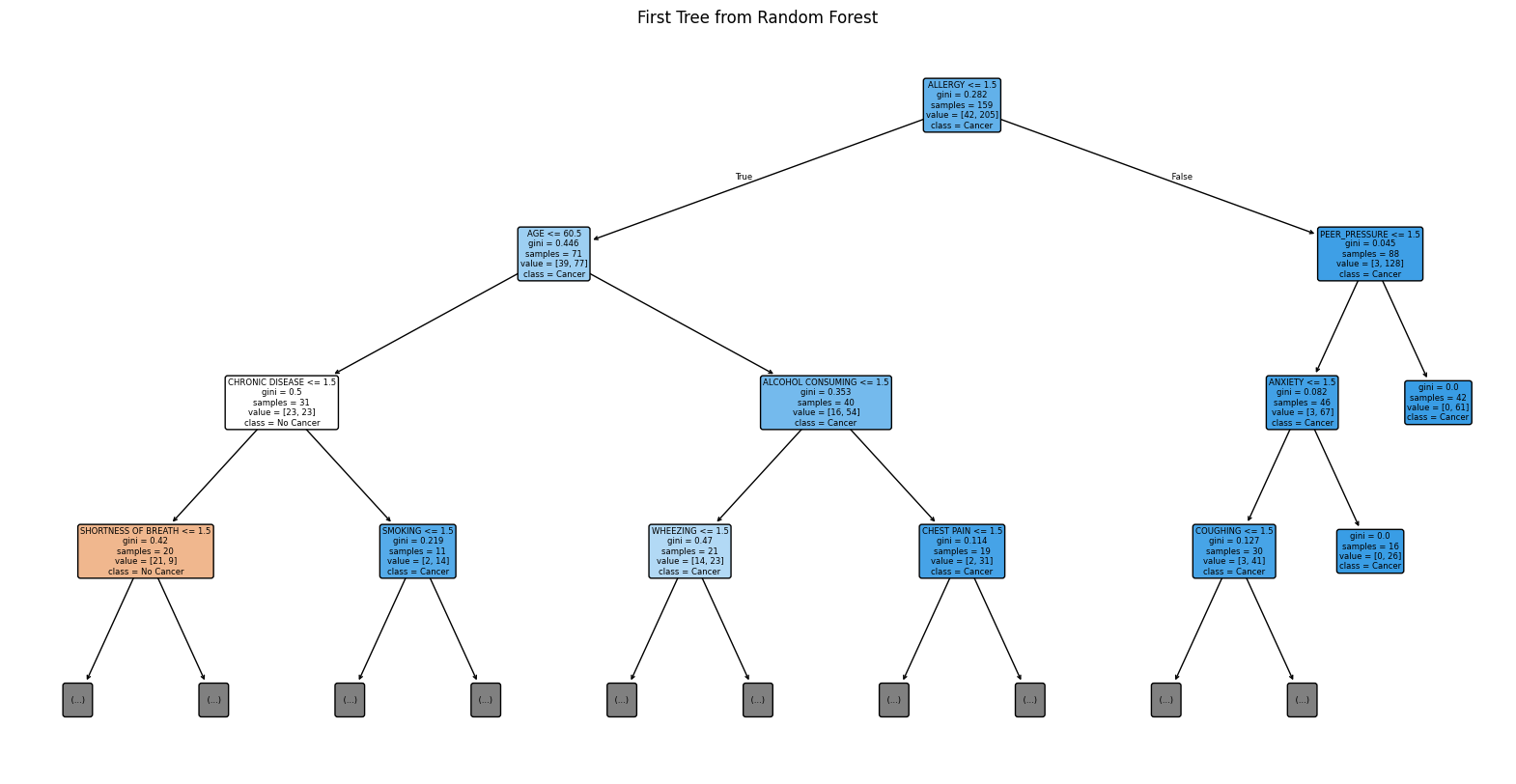


**Random Forest**

Accuracy: 96.7741935483871

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| No Cancer | 0.50 | 0.50 | 0.50 | 2 |
| Cancer | 0.98 | 0.98 | 0.98 | 60 |
| Accuracy |  |  | 0.97 | 62 |
| Macro Average | 0.74 | 0.74 | 0.74 | 62 |
| Weighted Average | 0.97 | 0.97 | 0.97 | 62 |





1. **Conclusion and Future Work**

Lung Cancer is a rapidly growing cancer. Each year, more and more people are detected by lung cancer. Patients and doctors are rapidly trying to find ways to detect lung cancer in early stage so that it can be treated. Lung Cancer is hard to cure it the later stages of cancer. This paper uses the basic life habits and information about the person to predict cancer. Many evaluation techniques like Mean Squared Error, Root Mean Square Error and R-Squared Score are used on different machine learning algorithms. Regression Models used to predict cancer are Linear Regression, Ridge Regression and Lasso Regression. Classifier Models like Logistic Regression, Decision Trees Classifier, and Random Forests were proved to have more accuracy in predicting lung cancer. This helps the patients and doctors to detect lung cancer easily just by their habits and basic information.

For further research, a mobile application can be developed by the researchers to take the information and habits of the user on a daily basis and warn them if they are at a risk of lung cancer. This can be done by advanced machine learning algorithms [19][20] for intelligently predicting cancer more accurately.

1. **References**

[1] World Health Organization (WHO) Report about Lung Cancer, [Who Report About Lung Cancer](https://www.who.int/news-room/fact-sheets/detail/lung-cancer)

[2] World Cancer Research Fund Report, [Lung cancer report | World Cancer Research Fund](https://www.wcrf.org/preventing-cancer/cancer-statistics/lung-cancer-statistics/page/2/)

[3] Benesty J, Chen J, Huang Y and Cohen I 2009 Pearson correlation coefficient. In Noise reduction in speech processing, 1-4. Springer, Berlin, Heidelberg.

[4] Benesty J, Chen J and Huang Y 2008 On the importance of the Pearson correlation coefficient in noise reduction. IEEE Transaction on Audio, Speech, and Language Processing, 16(4), 757-765.

[5] Towards Data Science, Linear Regression [Linear Regression Unlocks the Secrets of Complex Models | Towards Data Science](https://towardsdatascience.com/mastering-the-basics-how-linear-regression-unlocks-the-secrets-of-complex-models-8aa33920c105/)

[6] Wikipedia, Linear regression, [Linear regression, - Wikipedia](https://en.wikipedia.org/wiki/Linear_regression,)

[7] Montgomery D C, Peck E A and Vining G G 2012 Introduction to linear regression analysis, 821. John Wiley & Sons.

[8] Ridge Regression Data Camp, [Ridge Regression in Python | DataCamp](https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression)

[9] Lasso Regression Data Camp, [Lasso Regression in Python | DataCamp](https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression)

[10] Logistic Regression MIT Learning, [Logistic Regression ~ MIT-LEARNING](https://www.mit-learning.com/2024/10/logistic-regression-comprehensive-tutorial.html)

[11] Krishnaiah V, Narsimha G and Chandra N S 2013 Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. International Journal of Computer Science and Information Technology, 4, 39-45

[12] Tso G K and Yau K K 2007 Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9), 1761-68.

[13] Gini Impurity LearnDataSci, [Gini Impurity – LearnDataSci](https://www.learndatasci.com/glossary/gini-impurity/)

[14] Svetnick V, Liaw A, Tong C, Culberson J C, Sheridan R P and Feuston B P 2003 Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, 43(6), 1947-1958.

[15] Casanova R, Saldana S, Chew E Y, Danis R P, Greven C M and Ambrosius W T 2014 Application of random forests methods to diabetic retinopathy classification analyses. PLOS one.

[16] Masetic Z and Subasi A 2016 Congestive heart failure detection using random forest classifier. Computer Methods and Programs in Biomedicine, 130, 54-64.

[17] Lu S, Li Z, Qin Z, Yang X and Goh R S M 2017 A hybrid regression technique for house prices prediction. In 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 319-23. IEEE.

[18] Thompson P 1991 A Evaluation of the M-competition forecasts via log mean squared error ratio. International Journal of Forecasting, 7(3), 331-34.

[19] Ng A & Deisenroth M 2015 Machine learning for a London housing price prediction mobile application. Imperial College London.

[20] Aziz M A, Nurrahim F, Susanto P E and Windiatmoko Y 2020 Boarding Hose Renting Price Prediction Using Deep Neural Network Regression on Mobile Apps. arXiv preprint arXiv:2101.02033.