Lung Cancer Prediction

Srilekha Malraju (msrilekh@iu.edu), Rohith Venkata Reddy (rohi@iu.edu)

1. Abstract

Lung Cancer is one of the dreadful diseases prevailing in the world. Several deaths are being caused by Lung Cancer and there are several research studies performed on the cause factors. The American Cancer Society's estimates for lung cancer in the United States for 2022 are about 236,740 new cases of lung cancer (117,910 in men and 118,830 in women), of these about 130,180 deaths are caused from lung cancer (68,820 in men and 61,360 in women). Lung cancer is by far the leading cause of cancer death , making up almost 25% of all cancer deaths. For people who smoke the risk is much higher, while for those who don't, the risk is lower. Our focus in this project is to predict the Lung Cancer in the early stage that can be caused based on individual lifestyle, habits, and other factors. We have applied multiple machine learning techniques over the existing dataset to achieve best Accuracy score for correct prediction.

2. Keywords

Lung Cancer Prediction, Exploratory Data Analysis, Logistic/Linear Regression, Prediction Algorithms, Naïve Bayes, Random Forest, Support Vector Regression.

3. Introduction

Lung Cancer is one of the most frequent cause of deaths in the world. It can in multiple stages which can be divided into levels like low, medium, and high. While majority of the cases are attributed to smoking, exposure to air pollution is also a risk factor. The study, which was published in the journal Nature Medicine, looked at data from over 462,000 people in China who were followed for an average of six years. The participants were divided into two groups: those who lived in areas with high levels of air pollution and those who lived in areas with low levels of air pollution.

The researchers found that the people in the high-pollution group were more likely to develop lung cancer than those in the low-pollution group. They also found that the risk was higher in nonsmokers than smokers, and that the risk increased with age. Hence, we are using multiple factors in this dataset to identify the risk factors for lung cancer and predicting the likelihood of a patient developing lung cancer and its level in a patient.

We used multiple Machine Learning Regression models to predict the Level of Lung cancer in a patient based on multiple factors. We used Regression machine learning techniques to understand the relationship between Target variable (Level) and predictor variable relationships. Regression Model allows us to understand the levels observed in multiple independent variables are associated with levels of the Target variable. This helps us in determining the strength of factors that are responsible for Lung Cancer in a patient. We have also used Naïve Bayes methods in this project as these are expected to perform better for small datasets. This can easily solve multi-level prediction and predict the class of a test dataset.

4. About the dataset

The Lunger Cancer Prediction dataset is sourced from Kaggle. This dataset has 26 columns and 1000 rows. It contains information on patients with Lung Cancer, including their age, gender, air pollution exposure, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoker, chest pain, coughing of blood, fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of finger nails and snoring. Of these, Age is a numeric column and others are Categorical values which shows the level of a particular column in a patient. For example: Level of Smoking of the patient, Level of Alcohol use of the patient with Levels ranging from 1 to 8. The target variable Level has values Low, Medium and High. Patients data with Level of Cancer are equally distributed in this dataset for better analysis. About 37% of the patients have High Level of cancer, 33% with Medium Level and Rest 30% with Low level of cancer.

5. Methodology

The goal of this project is to predict the Lung Cancer in a patient based on certain factors. The methodology we followed is as below:

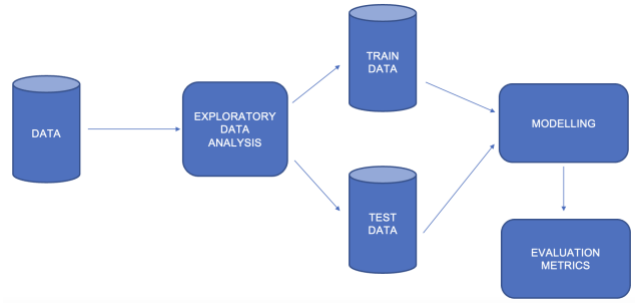

Figure 1: Project Methodology

Initially, we have considered the dataset and performed Exploratory Data Analysis where we have gained a lot of meaningful insights from the data. This helped us to identify most interesting variables and their correlation with the target variable. We have divided the dataset into Train and Test Dataset. We have built models like Logistic and Linear Regression, Naïve Bayes models, Random Forest and Support Vector Machine and performed Deep Learning Techniques. After that we split the data into test and training sets and trained them using Decision Tree classifier, K – Nearest Neighbors and Multi-Layer Perceptron. We also did some Hyperparameter Tuning and checked the performance of these models' using metrics like Accuracy, Precision and Recall. We have used evaluation metrics like Accuracy, F1-Score, mean square error, K-fold Cross Validation accuracies to find the best prediction model for this dataset.

6. Exploratory Data Analysis

We started by analyzing data to know what the interesting features are and how they are related with each other and with the target variable. To deduce that, we performed EDA on the dataset to understand the distributions of certain attributes and to understand the correlation between attributes. This helped us gain lot of meaningful insights from the data. Below are some of the results of EDA.

We have started with basic analysis like checking for null values, describing the data, and understanding the data types. Figure 2 shows the distribution of number of patients based on Cancer Level. We can see that around 350 patients have high level of cancer, 325 have medium level and 300 patients have low cancer level. Figure 3 shows the Gender Wise Age distribution. Here 1 refers to male and 2 refers to female. We could observe that between age 30 to 50 most of the cases are caused. After 50+, we could see most of them are male cases. This somewhere shows that in this dataset, cancer cause is independent of Age and gender as there are no significant observations.

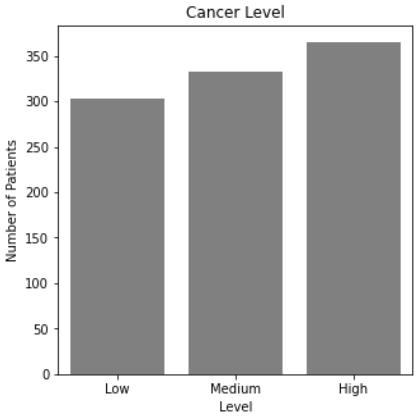

Figure 2: Cancer Level

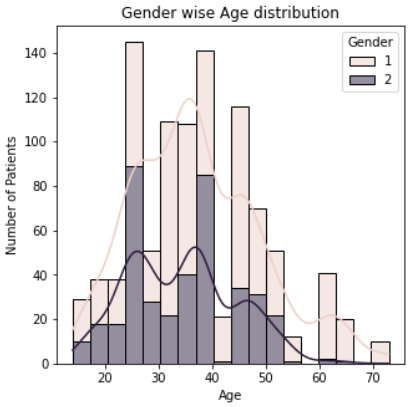

Figure 3: Gender wise Age distribution

Figure 4 shows the count sub plots for all the variables in single plot. We can observe certain variables like Air pollution, Dust Allergy, Genetic Risk, Smoking, Genetic Risk has high counts corresponding to high levels. We can understand variable levels in a patient in correspondence with cases counts. Figure 5 shows the positively correlated features and negatively correlated features with target. Figure 6 is the correlation heatmap and we can notice a small square of the highly correlated attributes which are the first 13 attributes between age and coughing of blood.

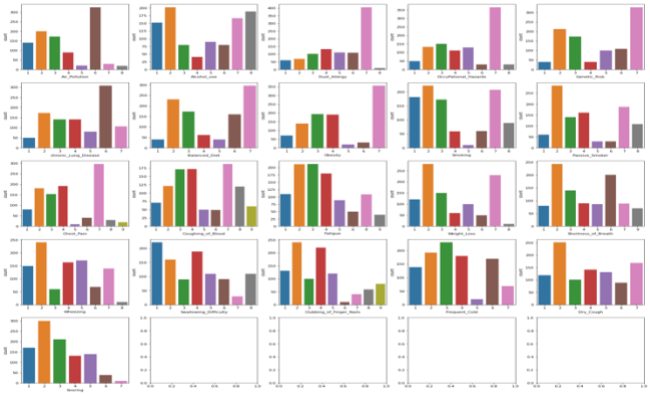

Figure 4: Count plots for all variables

```
Most Correlated features with Level of Lung Cancer:
 Passive Smoker       0.703594
Balanced Diet         0.706273
Dust Allergy          0.713839
Alcohol use           0.718710
Coughing of Blood     0.782092
Obesity               0.827435
Level                 1.000000
Name: Level, dtype: float64

Least Correlated features with Level of Lung Cancer:
 Gender                      -0.164985
Age                           0.060048
Wheezing                      0.242794
Swallowing Difficulty         0.249142
Clubbing of Finger Nails      0.280063
Snoring                       0.289366
Weight Loss                   0.352738
Name: Level, dtype: float64
```

Figure 5: Highly Correlated Variables with Target

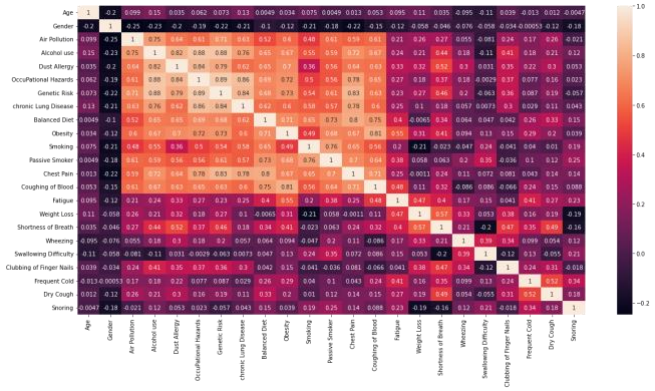

Figure 6: Correlation Heatmap

7. Modelling

A machine learning model will be trained on certain data to recognize patterns and learn from the train dataset. We use test data to test the efficiency of the model using evaluation metrics. In this project we have divide dataset into 70% of Train Data and 30% of Test data.

We have used several Regression models and Naïve Bayes methods described below:

Linear Regression: This is the basic regression model which attempts to use multiple independent variables to predict an outcome and a single continuous dependent variable. The equation for the Linear Regression Model can be defined as

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots\ldots\ldots + a_nX_n$$

where $X_1, X_2, \ldots..,X_n$ represents independent variables and the co-efficient represent the weights. This is a simple model yet doesn't not well for this dataset due to over-fitting.

Logistic Regression: It is a predictive algorithm using independent variables to predict the dependent variable just like Linear Regression, but with the difference that independent variable should be categorical variable. This is a statistical model which uses Logistic function to model the conditional probability. P is probability that event Y occurs. P/(1-P) is the odds ratio. Theta is the parameters of length m.

$$\ln\left(\frac{P}{1-P}\right) = \theta_1 + \theta_2 x + \mathrm{e}$$

$$\frac{P}{1-P} = \mathrm{e}^{\theta_1+\theta_2 x+\mathrm{e}}$$

$$P = \frac{1}{1+\mathrm{e}^{-}}(\theta_1 + \theta_2 x)$$

$$\sigma(z) = \frac{1}{1+\mathrm{e}^{-z}} \quad \text{Where} \quad z = \theta^T x$$

$$\theta^T \mathbf{x} = \sum_{i=1}^{m} \theta_i x_i = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

Ridge Regression: Ridge Regression is almost identical to Linear Regression except that we introduce a small amount of bias. In return for said bias, we get a significant drop in variance. In other words, by starting out with a slightly worse fit, Ridge Regression performs better against data that doesn't exactly follow the same pattern as the data the model was trained on. In Ridge Regression, the loss function is the linear least squares function and the regularization is given by the l2-norm.

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N}\left\{y_i - \sum_{j=0}^{M} w_j\, x_{ij}\right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$

Gaussian Naïve Bayes: Bayes' rule provides us with the formula for the probability of Y given some feature X. When there are multiple X variables, we simply use the formula

$$P(Y = k|X1, X2 \ldots . Xn) = \frac{\mathrm{P(Y)} \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P(Xi|Y)}}{P(X1) * P(X2) \ldots . * P(Xn)}$$

Gaussian Naïve Bayes is used when we assume all the continuous variables associated with each feature to be distributed according to Gaussian Distribution. The data we have mostly does not consists continuous values and hence doesn't follow this distribution. X and Y are the probabilities of the events.

Bernoulli Naïve Bayes: Naïve Bayes concepts is same as explained above but Bernoulli Naïve Bayes is used when input features are present only in binary form. It considers a Bernoulli distribution of a random variable X. The data we have is mostly categorical with different level values.

$$P(X) = \begin{cases} p & if\ X = 1 \\ q & if\ X = 0 \end{cases}$$

$$where\ q = 1 - p\ and\ 0 < p < 1$$

Random Forest Regression: A decision tree is a type of flowchart that shows a clear path to a decision. A decision tree starts with a master node, then branches in two or more ways. Each branch has its own set of probable outcomes, which include a variety of possibilities and random events until a conclusion is achieved. Looks like a tree structure. Random Forest Regression algorithms uses the combination of multiple random decision trees each trained on a subset of data. The use of multiple trees gives stability to the algorithm and reduces variance. The random forest regression algorithm is a commonly used model due to its ability to work well for most kinds of data. It can handle linear and non-linear relationships well and are not influenced by outliers to a fair degree. This algorithm can be used for both classification and regression tasks.

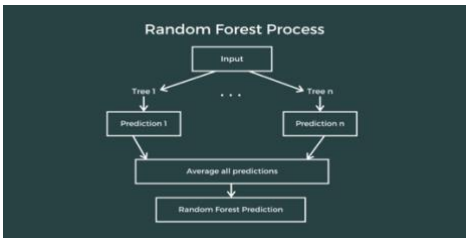

Figure 7: Random Forest Process

Support Vector Regression: SVM is a supervised machine learning algorithm that can be used for both classification or regression challenges. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. This method is robust to outliers and have excellent generalization capacity, with high prediction accuracy.

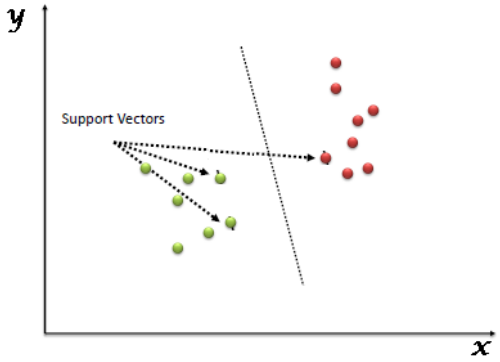

Figure 8: Vectors Space

8. Results and discussion

In order to evaluate the models we have built, we have used multiple metrics like Accuracy, Mean square error, F-1 Score, Precision, Recall, K-Fold Cross validation accuracies. Below are the definitions for the metrics.

Test Accuracy: Accuracy is number of correctly predicted data points out of all the data points. It can be described using the formula.

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN)$$

Mean Square Error: MSE measures the amount of error in the models. It is defined as average squared difference between observed and the predicted values. The model with less error produces more precise predictions.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Precision: Precision talks about how precise/accurate model is out of those predicted positive, how many of them are actual positive. This is good measure to determine when cost of False positive is more.

$$\text{Precision} = TP / (TP + FP)$$

Recall: Recall calculates how many of the Actual Positives our model capture through labeling it as True Positive. This metric is useful when there is high cost associated with False Negative.

$$\text{Recall} = TP / (TP + FN)$$

F1 Score: This is a function of both Precision and Recall. If we want to seek a balance between Precision and Recall and there are large number of Actual Negatives, this metric is used.

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

K-Fold Cross Validation accuracies: In this method, dataset is randomly split into k groups and each group is hold out as test data set and remaining are considered as training data set. Model is fit on the training set and then evaluate it using the test dataset. The test accuracies reported from all k folds are considered and 3 values are computed. Min Accuracy, Mean Accuracy and Max Accuracy from all the k-folds accuracies.

With the understanding of these metrics, we have computed all the metrics for our models in this project. Below are the results:

Model	Min Accuracy	Mean Accuracy	Max Accuracy	MSE	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.942857	0.977143	1.000000	0.020000	0.980000	0.980020	0.981340	0.980431
LinearRegression	0.881773	0.928326	0.948382	0.053333	0.946667	0.957672	0.943860	0.947571
GaussianNB	0.857143	0.900000	0.971429	0.176667	0.873333	0.887858	0.877671	0.876326
BernoulliNB	0.828571	0.894286	0.942857	0.170000	0.880000	0.894029	0.886124	0.881581
RandomForestRegressor	0.999932	0.999979	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000
SVR	0.987657	0.989094	0.991016	0.000000	1.000000	1.000000	1.000000	1.000000
SGDRegressor	0.867935	0.919498	0.936792	0.043333	0.956667	0.964770	0.954386	0.956920
Ridge	0.881583	0.928318	0.948468	0.053333	0.946667	0.957672	0.943860	0.947571

Figure 9: Results Table

We can observe that Accuracies, and all other metrics for Gaussian Naïve Bayes and Bernoulli Naïve Bayes are not good compared to other models. These are least performed models out of all we have built in this project. As this is a small dataset, we are focusing on K-Fold cross validation accuracies calculated with k=10 for better evaluation of the models. We could notice that the accuracies are high for Logistic Regression, Support Vector and Random Forest Regression. While Logistic Regression has good scores for all the metrics with Accuracy of 98%, Support Vector and Random Forest performs best with 100% Accuracy. To further verify these models, we can observe the K-fold accuracies. The mean accuracy with K-fold method is 99.99% for Random Forest Regressor and 98.9% for Support Vector Regressor. Hence, for this dataset, we can say that Support Vector Machines and Random Forest Regression are the best models.

9. Conclusion and future work

Lung Cancer is one of the diseases that causes high number of deaths in the world. Using Machine Learning, we predicted the Lung Cancer in a patient based on his habits and other factors. We have used Models like Logistic Regression, SVM, Naïve Bayes, Random Forest. By evaluating the model, we found that the best models are Random Forest and SVM for this dataset. We intend to use these models in the Real-world situations to predict the lung cancer situation in a patient in the early stage and make world a better place to live. Also, we agree that the analysis is performed on a small dataset, hence as a future work we would like to collect more data from many real-world sources, labs and hospitals across the globe and develop the machine learning model to achieve highest accuracy.

10. References

1. B. S, P. R and A. B, "Lung Cancer Detection using Machine Learning," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 539-543, doi: 10.1109/ICAAIC53929.2022.9793061.

2. C. Thallam, A. Peruboyina, S. S. T. Raju and N. Sampath, "Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1285-1292, doi: 10.1109/ICECA49313.2020.9297576.

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

4. D. Rawat, "Validating and Strengthen the Prediction Performance Using Machine Learning Models and Operational Research for Lung Cancer," 2022 IEEE International Conference on Data Science and Information System (ICDSIS), 2022, pp. 1-5, doi: 10.1109/ICDSIS55133.2022.9915898.

5. D. Reddy, E. N. Hemanth Kumar, D. Reddy and M. P, "Integrated Machine Learning Model for Prediction of Lung Cancer Stages from Textual data using Ensemble Method," 2019 1st International Conference on Advances in Information Technology (ICAIT), 2019, pp. 353-357, doi: 10.1109/ICAIT47043.2019.8987295.

6. Anil Kumar C, Harish S, Ravi P, Svn M, Kumar BPP, Mohanavel V, Alyami NM, Priya SS, Asfaw AK. Lung Cancer Prediction from Text Datasets Using Machine Learning. Biomed Res Int. 2022 Jul 14;2022:6254177. doi: 10.1155/2022/6254177. PMID: 35872862; PMCID: PMC9303121.

7. Tan, Steinbach, Kumar, Karpatne, "Introduction to Data Mining 2nd Edition"

8. https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html
9. https://deepai.org/machine-learning-glossary-and-terms/accuracy-error-rate
10. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
11. https://machinelearningmastery.com/k-fold-cross-validation/
12. https://www.tutorialride.com/data-mining/regression-in-data-mining.htm
13. https://towardsdatascience.com/quick-and-easy-explanation-of-logistics-regression-709df5cc3f1e
14. https://towardsdatascience.com/ridge-regression-python-example-f015345d936b
15. https://medium.com/analytics-vidhya/naive-bayes-algorithm-implementation-from-scratch-f9a2a12789b5
16. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
17. https://www.theclickreader.com/random-forest-regression/