Lung Cancer Risk Factors

Abstract—This paper contains findings of prediction made to diagnose lung cancer. Logistic Regression is utilized as machine learning methodology. This dataset consists of the dependent variable "lung cancer" and 15 potential risk factors which are correlated to lung cancer. The model has been assessed, explaining why it was specifically used and then present the final results. In this report data management and data cleansing is also discussed as well as exploratory data analysis (EDA) of the dataset. We will also show the process of reaching conclusions based on inference and analysis. A literature review of related works is also presented in this report to expand the scope of discussion for the reader. An evaluation of Logistic Regression is conducted to identify the accuracy of the model. Traditionally diagnosing lung cancer was based on mostly qualitative study of patients with doctors however the recent advancements in Machine Learning and Data Science techniques can now give a quantitative value to the risk or likelihood of someone being diagnosed with heart disease. Logistic Regression is a classification technique in supervised learning. The findings are presented in tabular form as well as visuals to aid the reader.

Keywords: smoking, risk factors, lung cancer, regression, prediction, prevention

1 INTRODUCTION

This project will investigate the prediction of whether lung cancer is caused by Anxiety, Smoking, Yellow Fingers, Peer pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol, Coughing, Shortness of Breath, Swallowing Difficulty, and Chest Pain using logistic regression model. Logistic regression model will take data, and create a prediction of whether a person will or won't have cancer and which factors are significant. Cancer is a real-world problem, and this project attempts to create a footstone for data, which could be potentially used for analysis of future trends. Men are far more likely to get cancer than women. As with the incidence figures, when translated into European age-standardised rates, the contrast between the sexes is more marked; the death rates in 2010 were 201.6 per 100,000 in males and 146.8 per 100,000 in females, respectively (mortality rate ratio equals 1.37 or 37% higher risk of death from cancer for men). [1]

Incidence rates are strongly related to age for all cancers combined, with the highest incidence rates being in older people. In the UK in 2016-2018, on average each year more than a third (36%) of new cases were in people aged 75 and over.

Age-specific incidence rates rise steeply from around age 55-59. The highest rates are in the 85 to 89 age group for females and males.

Incidence rates are significantly higher in females than males in the younger age groups and significantly lower in females than males in the older age groups. The gap is widest at age 40 to 44, when the age-specific incidence rate is 2.1 times higher in females than males. [2]

Cancer

Cancer is a disease in which some of the body's cells can grow without control and spread to other parts of the body. It can start almost anywhere in the human body. Normally, human cells grow and multiply to form new cells as the body needs them. When cells grow old or become damaged, they die, and new cells can take their place. This leads to the spread of tumours which then begin to cause life threatening serious conditions.

Every year over 250,000 people in England are diagnosed with cancer, and over 130,000 people die from this disease. This costs the NHS massive amounts, over £5 billion, yet the cost to the whole of society; given the loss of productivity and life, is estimated to be around £18.3billion. [3]

2 OBJECTIVES

- Identify correlation, between risk factors and lung cancer.
- Identify heavily weighted risk factors that contribute lung cancer.
- Conduct EDA to find correlation between certain risk factors in the dataset that have a higher relation to each other in determining lung cancer or not using ML.
- Utilize ROC curve to visually demonstrate accuracy.
- Analyze strengths and weaknesses of the Logistic Regression model and evaluate.

3 LITERATURE REVIEW

In this section of the report, related discussions and a literature review will be presented and dissected. The work discussed will be evaluated in relation to the research problem being investigated.

The first report to be discussed is **"Smoking and lung cancer"** [4] by *Tevfik Ozlü, and Yilmaz Bülbül* the study focussed on smoking and its relation with lung cancer, as around one-third of adults are known to be smokers, and smoking rates are increasing among the female population. The association between cigarettes and lung cancer has been proven by large cohort studies. Tobacco use has been reported to be the main cause of 90% of male and 79% of female lung cancers. 90% of deaths from lung cancer are estimated to be due to smoking. The risk of lung cancer development is 20-40 times higher in lifelong smokers compared to non-smokers.

The second report that will be discussed as part of the literature review is titled "Analysis of conditional logistic regression for risk factors of lung cancer in Dachang Tin Mine" by *K G Wu* [5] An increased risk of lung cancer in Dachang Tin Mine of Guangxi has been reported. To investigate the factors of the excessive risk of lung cancer, the authors conducted a matched pair case-control study in the mine area and analysed the effect of multiple factors, such as condition of living and housing, occupational exposure and smoking by statistical method of conditional logistic regression. The patients group consisted of 69 patients with primary bronchogenic cancer including 55 deceased. The control group consisted of 138 persons also including 55 deceased. The results showed that the factors of the excessive risk of lung cancer in the mine area were mainly related to the occupational exposure. The risk factors with statistical significance in conditional logistic regression analysis were exposure time of smelting, time of underground drilling, and age of beginning mining underground.

4 DATA MANAGEMENT

4.1 Data Source

The dataset was found from Kaggle, analysing whether Anxiety, Smoking, Yellow Fingers, Peer pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol, Coughing, Shortness of Breath, Swallowing Difficulty, and Chest Pain cause lung cancer, The dataset has 15 attributes along with a target attribute, giving us a binary option of whether the patient has lung cancer. The shape of the dataset is 303x16. The sex of person is the string M for male and F for female, and the rest of the options present 2 for yes and 1 for no of whether they inhabit those features.

4.2 Dealing With Duplicates

Sometimes in a dataset we have a problem with duplicated data values, in which some data were duplicated more than one time. Duplicated date can have a significant effect on the conclusions one established from the dataset, hence duplicated data can present a large issue.

- Duplicated Values: 33 values were duplicated in the dataset so, we removed them.

Table 1: Data Sample from original dataset

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY
M	69	1	2	2	1	1	2	1
M	74	2	1	1	1	2	2	2
F	59	1	1	1	2	1	2	1
M	63	2	2	2	1	1	1	1
F	63	1	2	1	1	1	1	1

Table 2: Data Sample from original dataset

SURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CH P
1	1	2	1	2	2	2	2	2	
1	2	2	2	1	1	1	2	2	
2	1	2	1	2	1	2	2	1	
1	1	1	1	1	2	1	1	2	
1	1	1	1	2	1	2	2	1	

4.3 External Libraries

1. **NumPy:** A python library that includes support for extremely large multidimensional arrays and matrices. It is used regularly due to the fact it provides access to a large library of high-level mathematical functions and operations on these arrays and matrices.
2. **Pandas:** A Python data framework library, a regular choice for pythonic programmers due to its ability to manipulate and present data. A Data Frame is by definition a 2d data matrix that supports many operations on it for convenience.
3. **Matplotlib:** A Python library utilised to visualise data, with an abundance of graphs and operations which can be developed within this library. The graphs' axis and scales can be changed accordingly and some elements of customisation in colour are possible.
4. **Scikit-learn:** A Python machine-learning library, that is used regularly within Data Science and computing. It's implementation of algorithms, were needed in order to use Logistic Regression.

5 METHODOLOGY

The approach we take in diagnosis and prediction of heart disease is we use Logistic Regression model, this is a classification model using the data features "Risk Factors" to provide a forecast to whether a patient has or does not have Lung Cancer,

explanation and reasoning behind the model is explained as follows:

Logistic Regression Algorithm

Logistic regression is an optimization method that is based on boundaries in classifier models. A boundary is set usually between 0-1, when the weight and vectors are multiplied the outcome will correspond with the distance from the boundary. Whether the outcome is larger/smaller than the boundary will determine which section this particular instance of the predictor space is in, in our case our predictor space will be whether a patient HAS/HAS NOT lung function. We use this model as a classifier and not as an optimizer. A P(X) value is given as a result of the logistic function determining the probability of an instance belonging to a section in the prediction space, 1-P(X) is the probability of that same instance belonging in the opposite categorical section.

Logistic Regression Justification

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". [6]

In a binary logistic regression model, the dependent variable has two levels (Diseased, and Not Diseased). Outputs with more than two values are modelled by multinomial logistic regression and, if the multiple categories are ordered, by ordinal logistic regression (for example the proportional odds ordinal logistic model). [7]

6 Exploratory Data Analysis

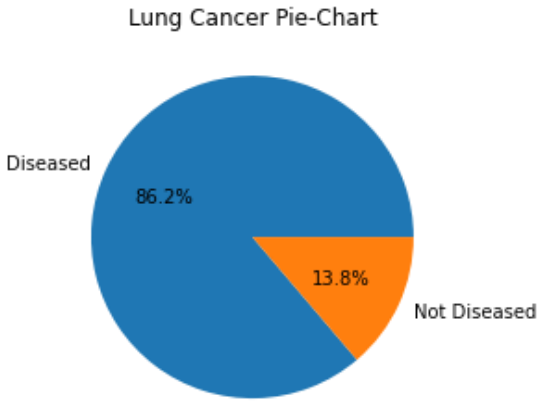
In order to understand every factor and the prevalence of interrelationships between contributing factors to lung cancer, exploratory data analysis will need to be performed upon the dataset.

From the Fig. 2 it is clear that the dataset is fairly even in terms of population, 52.4% are male, and 47.6% are female. This shows that we can observe a clear trend in regards to gender as a factor, and see if the observation that men are more likely to get cancer is true. To continue the analysis, we will now look in the distribution of the population in terms of age. The population is fairly evenly distributed in terms of age; however, it is more concentrated in the ages between 55-65, which is a common age range to be diagnosed with cancer. Furthermore, there seems to be far more males, per year of age, with the exception of a couple anomalies, namely at the age of 60 and 55, where there are far more women than men.

Lung Cancer

We will use pie-plot from matplotlib to visualise our findings and it will give us an idea how much percentage of population is Diseased and Not Diseased. As from the pie chart we found out that 86.2% of the population is Diseased that is colour blue and 13.8% of the population is Not Diseased.

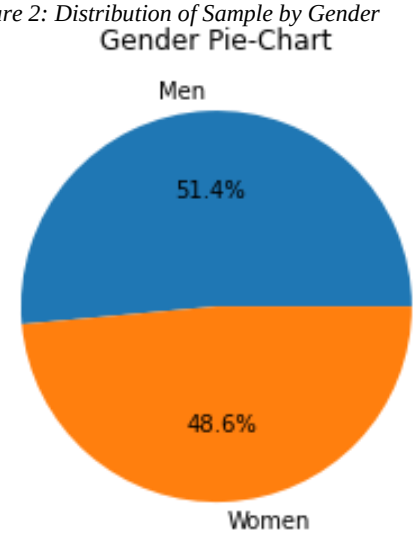
Figure 1: Distribution of Sample by Lung Cancer



Gender

As from the pie chart we found out that 51.4% of the population is Men that is colour blue and 48.6% of the population is Females.

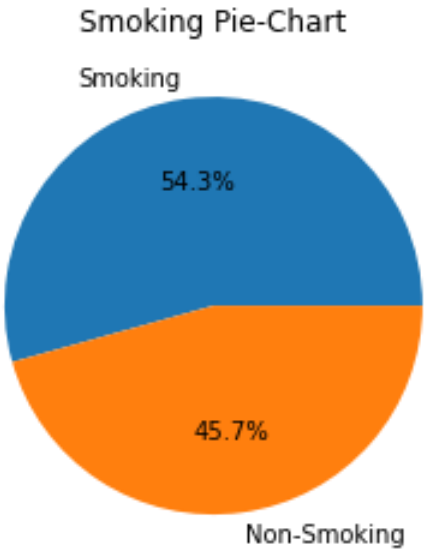
Figure 2: Distribution of Sample by Gender



Smoking

As from the pie chart we found out that 54.3% of the population Smoke that is colour blue and 45.7% of the population doesn't smoke.

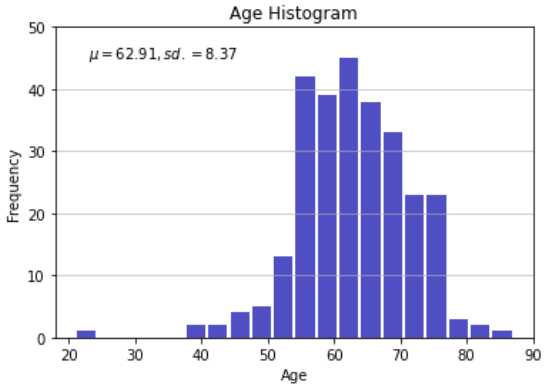
Figure 3: Distribution of Sample by Smoking



Age

As from the Age histogram we found out that it looks normal "Bell Curve" but a little skewed to the left however, it is more concentrated in the ages between 55-65 with Mean= 62.91 years, and Std. = 8.37 years.

Figure 4: Distribution of Sample by Age



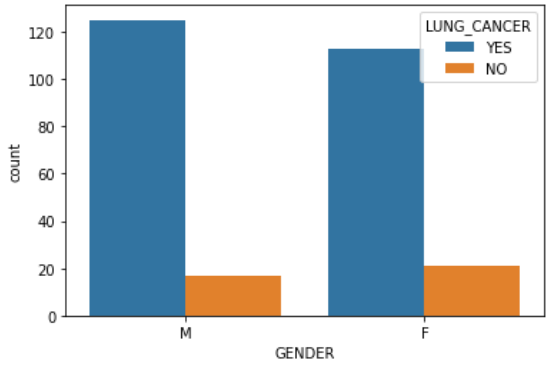
Cross Tabulation

According to various sources, the chances of getting cancer increases by age and by sex. In order to understand the prevalence of each factor in the dataset source, we need to identify them. In order to understand we need to use the pandas value _counts() function.

1. Lung Cancer Vs. Gender

As from the bar chart we found out that Males are more likely to get lung cancer than women.

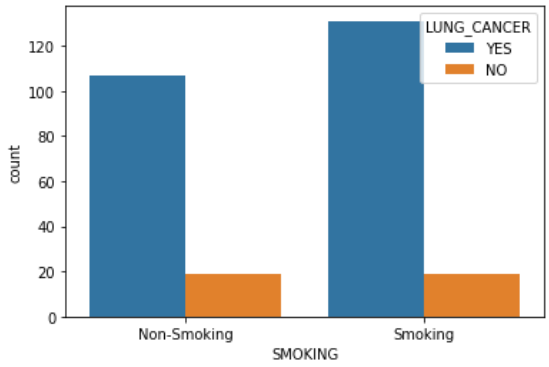
Figure 5:Lung Cancer Vs. Gender



2. Lung Cancer Vs. Smoking

As from the bar chart we found out that Smokers are more likely to get lung cancer than Non-Smokers.

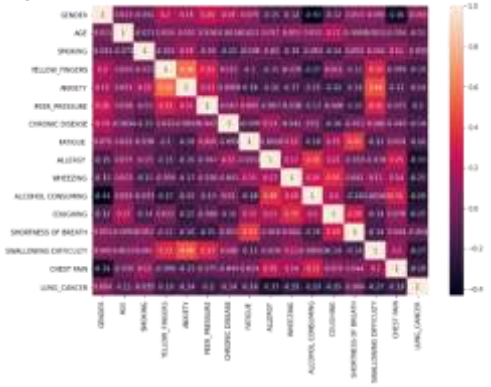
Figure 6: Lung Cancer Vs. Smoking



Correlation Matrix

In this stage of the Exploratory Analysis, we find the correlation matrix between all the risk factors and the dependent variable **"Lung Cancer".**

Figure 7: Correlation Matrix



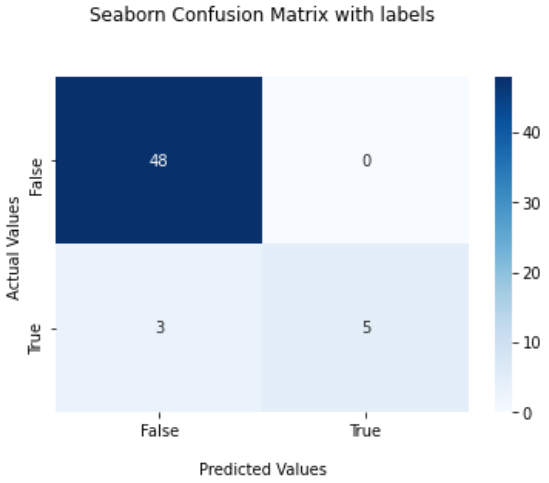
7 Logistic Regression

Now we can move onto our classifier model we used on the dataset Logistic Regression. This model is from same sklearn library and is practically applied in a similar fashion. from sklearn. linear_model import LogisticRegression The results will be more restricted to smaller output of either yes or no, in this case it will be either 0 or 1, this has been previously explained in the methodology section. The logistic regression method is simpler to apply, and the method used to classify the data is straight, multiplying the weight and predictors to plot an output between the given boundary as explained in more detail in previous sections. The process goes as: split the data between test and train in this case being 80% train and 20% test.

Confusion Matrix

This is the resulting confusion matrix and accuracy; the accuracy is 0.95 which is indicator of higher level of accuracy.

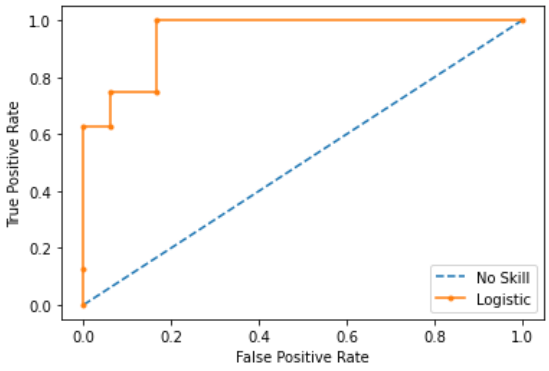
Figure 8: Logistic Regression Confusion Matrix



ROC Curve

We also have a ROC plot to show the rate at which the model performs when you move the boundary of the classifier. This produces an AUC score of 95%+ which is a reflection of a good model, considering the high dimensions of the dataset and the complexity of the problem.

Figure 9: ROC Curve



Logistic Regression Model Summary

The following summary shows that the model is significant (P-value < 0.05), and Pseudo R-square = 0.3666 which ranges between 0.2 – 0.4 indicating excellent model fit.

Figure 10: Logistic Regression Summary

Logit Regression Results

Dep. Variable:	LUNG_CANCER	No. Observations:	220
Model:	Logit	Df Residuals:	205
Method:	MLE	Df Model:	14
Date:	Mon, 28 Mar 2022	Pseudo R-squ.:	0.3666
Time:	09:02:35	Log-Likelihood:	-55.502
converged:	True	LL-Null:	-87.628
Covariance Type:	nonrobust	LLR p-value:	2.081e-08

Logistic Regression Coefficients

The following table shows logistic regression coefficients and Significance level, we can notice that Age, Chronic Disease, Fatigue, Allergy, Couching, Shortness of Breath, and Swallowing Difficulty.

While Gender, Smoking, Yellow Fingers, Anxiety, Peer Pressure, Wheezing, Alcohol Consuming, Chest Pain have no significant effect on Lung cancer.

Table 3: Logistic Model Coefficient

	coef	std err	z	P>|z|
GENDER	-0.5765	0.613	-0.941	0.347
AGE	0.1781	0.043	4.121	0.000
SMOKING	-0.8027	0.569	-1.411	0.158
YELLOW_FINGERS	-0.3480	0.657	-0.530	0.596
ANXIETY	0.4540	0.737	0.616	0.538
PEER_PRESSURE	-0.9489	0.563	-1.687	0.092
CHRONIC DISEASE	-1.5200	0.665	-2.292	0.022
FATIGUE	-1.7507	0.595	-2.941	0.003
ALLERGY	-1.4722	0.682	-2.159	0.031
WHEEZING	-0.5533	0.617	-0.897	0.370
ALCOHOL CONSUMING	-1.2038	0.630	-1.911	0.056
COUGHING	-2.0021	0.797	-2.511	0.012
SHORTNESS OF BREATH	1.4057	0.679	2.070	0.038
SWALLOWING DIFFICULTY	-1.5518	0.700	-2.218	0.027
CHEST PAIN	0.5409	0.570	0.964	0.335

8 Conclusion

We believe that the aims of the project have been achieved, we set out with objective aims and they

have been identified in the report along with evidence. One of the first aims was to the be able to diagnose and therefore predict lung cancer disease given the patient data in the form of the cancer.CSV file. We used logistic regression model that output result with relatively high accuracy, and provided a strong AUC number in the ROC, so we believe this aim was achieved. Also, we succussed in identifying the significant risk factors which affect the lung cancer.

9 REFERENCES

G., W. K. (1989). *Zhonghua zhong liu za zhi.* Chinese journal of oncology.

GOV.UK. (2022, 03 28). *1*. Retrieved from 1: https://www.gov.uk/government/publications/2010-to-2015-government-policy-cancer-research-and-treatment/2010-to-2015-government-policy-cancer-research-and-treatment

Ozlü, T., & Bülbül, Y. (2005). *Smoking and lung cancer.* Tuberkuloz ve toraks.

Sung, H, Ferlay, J, Siegel, RL, Laversanne, M, Soerjomataram, I, Jemal, A, Bray, F. (2020). *GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin. doi:https://doi.org/10.3322/caac.21660

Tolles, Juliana; Meurer, William J. (2016). *Logistic Regression Relating Patient Characteristics to Outcomes.* JAMA. doi:doi:10.1001/jama.2016.7653. ISSN 0098-7484

UK, C. R. (2022, 03 28). *1*. Retrieved from 1: https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/age#heading-Zero

Walker, SH; Duncan, DB. (1967). *Estimation of the probability of an event as a function of several independent variables.* Biometrika. doi:doi:10.2307/2333860. JSTOR 2333860