Lung Cancer Prediction

Ashar Shamim 21616

Muhammad Abdul Rafay 21563

Muhammad Awais 22183

Omer Azmi 21663

Group 26

Institute of Business Administration

Spring 2023

Submitted to:

Hassaan Khalid

Analytical Approach to Marketing Decisions

Table of Contents

1 Executive Summary

This project aims to develop a machine-learning model for the early prediction of lung cancer using a comprehensive dataset. Lung cancer is a significant health concern, and early detection plays a crucial role in improving treatment outcomes and reducing mortality rates. By analyzing an extensive dataset comprising patient records, this project utilizes machine learning techniques to accurately predict lung cancer risk factors. The developed models offer a quick and efficient method for computing the probability of lung cancer, enabling timely interventions and personalized healthcare. The findings from this project have the potential to assist medical professionals in identifying high-risk individuals and implementing targeted prevention and treatment strategies. The success of this project demonstrates the promising application of machine learning such as multinomial logistic regression and decision tree analysis in improving lung cancer prediction and ultimately contributing to better patient care.

2 Data Structure

The dataset used in this project consists of 1000 patient records obtained from an international dataset. The dataset contains 26 variables, including two identifier variables: Index and patient ID. These variables help uniquely identify and organize the data. The dataset was sourced from Kaggle, a popular online platform for data science and machine learning competitions.

For more information on the dataset refer to *link to the dataset.*

For more information on the study refer to *link to the Study.*

3 Feature Engineering

As part of the feature engineering process, we made two key transformations on the dataset.

Firstly, we converted the target y variable 'Level' from the original categorical labels of 'Low,' 'Medium,' and 'High' to numerical values of 1, 2, and 3, respectively.

Secondly, we transformed the 'Gender' variable, initially represented as '1' for Male and '2' for Female, to a binary representation of '0' for Male and '1' for Female. This conversion ensures compatibility with models that expect binary inputs and simplifies the interpretation of the gender variable as a binary attribute.

4 Problem Statement

Early prediction of lung cancer is crucial for improving treatment outcomes and reducing mortality rates. However, existing methods often require costly hospital tests and extensive data input, making them inaccessible and time-consuming. The tool utilizes machine learning techniques and analyzes a dataset of 1000 patient records to compute the risk factors associated with lung cancer. The objective is to provide an accurate and efficient method for computing the risk factors, enabling individuals who lack access to costly tests to obtain their lung cancer risk assessment quickly. The report concludes with an analysis of the tool's predictions and its supposed applications/recommendations.

5 Multinomial Logistic Model.

5.1 Loading the data

We loaded the final file into R using the *read.csv* command.

Our y variable is a categorical variable so we converted its data type into factor using the *as.factor()* command.

5.2 Splitting the data into training and test

We split the data into training and testing using a ratio of 7:3, using the code below.

```
#Data partition into training and testing
set.seed(123)
ind <- sample(2, nrow(mydata),
              replace = TRUE,
              prob = c(0.7,0.3))
training <- mydata[ind==1,]
testing <- mydata[ind==2,]

```

5.3 Training the Model

Multinomial logistic regression uses the *mulitnom()* command in the *nnet* package.

5.3.1 First Model

In multinomial logistic regression, the process of releveling is done to set a reference category for the outcome variable. The reference category is the baseline or comparison category against which the other categories are compared.

By releveling, we changed the reference category to "1" i.e., "Low" and thus interpret the model coefficients relative to this chosen reference. We chose 1 or Low as the reference level since it is considered a default value for a healthy person. It allowed us to compare the other categories against the reference category and understand their relationship in terms of the odds ratios or probabilities.

```
15  #multinomial logistic regression on training data
16
17  library(nnet)
18
19  training$Level <- relevel(training$Level, ref="1")
20  mymodel <- multinom(Level~.,
21                      data = training)
```

5.3.1.1 Calculating the significance values

We then calculated the significance level of each variable as the *multinom* command doesn't calculate the significant values by default.

```
27  #calculating significance level
28  MASS::dropterm(mymodel, trace=FALSE, test="Chisq")
```

The result (Appendix A) shows that all variables used in the initial model were insignificant.

5.3.2 Second Model

We ran a stepwise command on the first model to keep only the variables which are contributing significantly to the model. This reduced the variables and only 8 remained from the initial 23 X variables used. A summary of the step model is given in (Appendix B).

5.3.2.1 Calculating the significance values

The results (Appendix C) shows that the variables remaining: alcohol.use, dust.allergy, occupational.hazards, genetic.risk, coughing.of.blood, fatigue, wheezing, and snoring were highly significant.

We are using the second model in our study because it only makes use of significant variables and they best explain the outcome variable.

5.3.3 Accuracy of the model

5.3.3.1 Confusion matrix of training data

Confusion matrix of the training data shows that there are no false positives and false negatives in the data, and the model was able to predict everything correctly.

```
>   #Confusion matrix & Mis classification error - Training Data
>   trainingpred <- predict(stepmodel, training)
>   Tab <- table(trainingpred,training$Level)
>   Tab

trainingpred   1   2   3
           1 214   0   0
           2   0 225   0
           3   0   0 266
```

5.3.3.2 Predictions on the testing data.

We then used the predict command on data the model has not seen before and developed a confusion matrix. The model was able to make 100% accurate predictions as there were no false positives and false negatives.

```
>   #Confusion matrix & Misclassification error - Testing Data
>   testingpred <- predict(stepmodel, testing)
>   Tab1 <- table(testingpred, testing$Level)
>   Tab1

testingpred   1   2   3
          1  89   0   0
          2   0 107   0
          3   0   0  99
```

6 Factor Analysis

Factor analysis was conducted using SPSS to identify the smaller number of underlying factors that account for most of the variation among the variables.

The CSV data file was first imported into SPSS and the y-variable (Level) was removed from the dataset. All the remaining variables were then entered into the factor analysis. Scree plot and rotated solution were included to better aid in understanding. Moreover, coefficients with absolute values below 0.40 were suppressed to clean the result and allow them to be analyzed easily.

KMO and Bartlett's test was also used to determine if the dataset is appropriate for conducting factor analysis. The KMO value of this dataset was 0.775, which was greater than the threshold of 0.6, indicating that the sample was adequate, and the data was well suited for factor analysis. Similarly, the p-value of the Bartlett's test was calculated as less than 0.001 (significant), indicating that the observed variables in the dataset were significantly correlated, again making them suitable for factor analysis. Both these values suggested that substantial correlation exists within the dataset.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.775
Bartlett's Test of Sphericity	Approx. Chi-Square	22409.824
	df	253
	Sig.	<.001

6.1 Results

The factor analysis (Appendix D) showed that 6 out of the 23 variables were capturing 77.62% of the total variation while the remaining 17 variables only captured 22.38% of the variation. By analyzing the rotated component matrix, we could classify these 6 factors, however, as no proper logic could be determined to name these factors, they were simply taken as Factor 1, Factor 2 and so on.

6.2 Reliability Testing

To see how reliably the 6 factors could be used as scales in the future, their reliability was tested through Cronbach's Alpha. The given variables within each factor were entered together and the reliability of each factor was calculated as given below.

Factor	**Cronbach's Alpha**
Factor 1	0.910
Factor 2	0.591 (before reverse coding) 0.626 (after reverse coding)
Factor 3	0.652
Factor 4	0.501
Factor 5	0.562
Factor 6	0.032

A column was also added to show what Cronbach's Alpha would be if specific variables were deleted from that scale. This helped in identifying the variables without which the scale would be more reliable. Lastly, for Factor 2, the snoring variable was reverse coded due to it having a negative correlation, after which the reliability of the scale improved.

6.3 Interpretation

Factor analysis states that there are 6 underlying factors (Appendix E) that contribute to the observed relationships among variables. These groups of variables are quite helpful in understanding which variables are highly correlated with one another and make further analysis much easier through lowering the complexities of the dataset by reducing the 23 variables to only 6.

These 6 factors could also be used as scales in the future depending on their reliability. However, as seen through the reliability test, only Cronbach's Alpha of Factor 1 was calculated to be greater than 0.7, and hence, only that scale can be considered as reliable for future use. The other factors have poor reliability and should not be used as scales in the future.

7 Decision Tree Analysis

7.1 Loading the data

We loaded the final file into R using the *read.csv* command.

Our y variable is a categorical variable so we converted its data type into factor using the *as.factor()* command.

7.2 Splitting the data into training and test

We split the data into training and testing using a ratio of 7:3, using the code below.

```
#Datapartition
set.seed(123)
ind <- sample(2, nrow(mydata),
              replace = TRUE,
              prob = c(0.7,0.3))
train <- mydata[ind==1,]
test <- mydata[ind==2,]
```

7.3 Training the Decision Tree Model

We trained the decision tree model using the rpart command in rpart package.

A decision tree model (model1) using the "Level" variable as the target and all other variables in the training set as predictors.

```
#Train a decision tree model
model1 <- rpart(formula = Level~., data = train)
model1
summary(model1)
rpart.plot(model1)
```

We then used the rpart.plot(model1) command to visualize the decision tree as a plot. (Appendix F)

7.4 Training the Decision Tree Model using a Minbucket

As seen in (Appendix F), some of the terminal nodes had only 1% or 2% of the data. The model would not be able to give satisfactory results because of limited data, hence we used the minbucket of 30 this time which means that the model will only split if there are at least 30 values in the node.

```
#Training a model with minbucket = 30
model2 <- rpart(formula = Level~., data = train,
                control = rpart.control(minbucket =30))
model2
rpart.plot(model2)
```

We then used the rpart.plot(model1) command to visualize the decision tree as a plot. (Appendix G)

7.4.1 Accuracy of the model

7.4.1.1 Confusion matrix and accuracy of training data

Created a confusion matrix (t) comparing the predicted values (trainpred) with the actual values (train$Level) in the training set.

sum(diag(t))/sum(t): Calculates the accuracy of the training set predictions by dividing the sum of correctly predicted instances (represented by the diagonal elements of the contingency table t) by the total number of instances (sum(t))

This model is found to be about 94% accurate on the training data set.

```
> # Predict on training data
> trainpred<- predict(model2, newdata = train, type = "class")
>
> #Confusion matrix of training data
> t <- table(prediction = trainpred, actual = train$Level)
> t
          actual
prediction   1   2   3
         1 194   0   0
         2  20 219  19
         3   0   6 247
>
> #Accuracy of training data
> sum(diag(t))/sum(t)
[1] 0.9361702
```

7.4.1.2 Confusion matrix and accuracy of testing data

Created a confusion matrix (t2) comparing the predicted values (testpred) with the actual values (test$Level) in the training set.

sum(diag(t2))/sum(t2): Calculates the accuracy of the testing set predictions by dividing the sum of correctly predicted instances (represented by the diagonal elements of the contingency table t2) by the total number of instances (sum(t2))

This model is found to be about 88% accurate on the testing data set.

```
> #Prediction of test data
> testpred <- predict(model2, newdata = test, type = "class")
>
> #Confusion matrix of testing data
> t2 <- table(prediction = testpred, actual = test$Level)
> t2
          actual
prediction  1   2   3
         1 68   0   0
         2 21 103  10
         3  0   4  89
>
> #Accuracy of testing data
> sum(diag(t2))/sum(t2)
[1] 0.8813559
```

8 Comparing Models

Decision tree was able to make predictions on the training and testing data with an accuracy of 94% and 88% respectively. Whereas the multinomial logistic regression was able to make predictions with 100% in both data sets.

Since, identifying the risk of cancer involves a matter of life and death, even a 1% in the accuracy of the model would prove to be beneficial. Hence, we believe that the multinomial model would be a better option.

However, 100% accuracy is rare occurrences, and such accuracy is raising alarms for some issues in the data that were left unidentified. Sir Hassan suggested that it might be due to a variable being directly correlated with the outcome variable. However, we found no such variable in our exploratory data analysis. (Appendix H)

9 Recommendations

1. **Cost Saving:** The most important recommendation is that the model's findings and predictions should be shared with healthcare providers and specialists involved in lung cancer diagnosis and treatment. Given that cancer tests are quite costly, they can use this as an alternative to the tests to predict who has lung cancer based on multiple underlying factors, without any cost, and with great accuracy. Thus, even those who do not have the financial means to afford multiple medical tests can, to a great extent, know if they have a risk of lung cancer or not. Even when not used as an alternative, this model can give them an additional tool to support their clinical decision-making process, helping them make better decisions.
2. **Early Identification of Risk and Personalized Treatment Plans:** The model's predictions should also be used to identify the risk in individuals. The high-risk individuals should be recommended for regular screening tests and to undergo medical

evaluation which can increase the chances of successful treatment for these individuals. The model should also be used to not only predict the probability of lung cancer but also to suggest personalized treatment plans to these individuals based on individual risk factors and health profiles. This could help healthcare providers to tailor their treatment according to everyone, increasing their chances of survival.

3. **Prevention Programs:** Even better than curing cancer is the prevention of cancer. Therefore, those who are at risk but have not yet developed lung cancer should be entered into programs to improve their lifestyle. This model shows that alcohol consumption, smoking and obesity are all significant factors that contribute to lung cancer. Hence, these individuals who are classified by the model as highly likely to get lung cancer should go through programs such as smoking cessation programs and alcohol rehab programs so that lung cancer could be prevented.
4. **Technology Integration:** Moreover, this model should be integrated with technology to make it more dynamic. For example, if the model's predictions are connected to electronic health record systems, healthcare providers can identify patients at high risk of developing lung cancer during routine visits and perform screenings and evaluations. Similarly, mobile applications and wearable devices should be used to collect real-time health data from individuals. This again would further help in monitoring lung cancer risk and in giving health recommendations.
5. **Basis for New Models:** Lastly, as this model does quite well to predict lung cancer, it should be used as a basis to develop other models that predict other types of cancer. The underlying factors may change, but the learnings of this model can be used to develop further models to predict breast cancer, skin cancer and so on.

10 Appendices

10.1 Appendix A

```
Model:
Level ~ Age + Gender + Air.Pollution + 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 +
    Frequent.Cold + Dry.Cough + Snoring
                         Df AIC         LRT Pr(Chi)
<none>                       96
Age                       2  92 -1.4938e-05       1
Gender                    2  92  3.7588e-05       1
Air.Pollution             2  92 -4.4462e-05       1
Alcohol.use               2  92 -3.6392e-05       1
Dust.Allergy              2  92 -1.0730e-06       1
OccuPational.Hazards      2  92 -2.1783e-05       1
Genetic.Risk              2  92  3.2100e-05       1
chronic.Lung.Disease      2  92  3.3375e-05       1
Balanced.Diet             2  92  1.2504e-05       1
Obesity                   2  92 -2.0773e-05       1
Smoking                   2  92 -2.0936e-05       1
Passive.Smoker            2  92  3.5330e-06       1
Chest.Pain                2  92 -2.6722e-05       1
Coughing.of.Blood         2  92 -2.4216e-05       1
Fatigue                   2  92  3.6260e-05       1
Weight.Loss               2  92  4.5596e-05       1
Shortness.of.Breath       2  92 -7.3360e-06       1
Wheezing                  2  92 -1.4709e-05       1
Swallowing.Difficulty     2  92  5.2466e-05       1
Clubbing.of.Finger.Nails  2  92 -5.4612e-05       1
Frequent.Cold             2  92 -2.5028e-05       1
Dry.Cough                 2  92  1.1386e-05       1
Snoring                   2  92 -2.6412e-05       1
```

10.2 Appendix B

```
> summary(stepmodel)
Call:
multinom(formula = Level ~ Alcohol.use + Dust.Allergy + OccuPational.Hazards +
    Genetic.Risk + Coughing.of.Blood + Fatigue + Wheezing + Snoring,
    data = training)

Coefficients:
  (Intercept) Alcohol.use Dust.Allergy OccuPational.Hazards Genetic.Risk
2   -1214.670   -43.90828     60.01283           -138.16542     212.5988
3   -1839.242   124.47223   -104.28326            -90.23723     161.1853
  Coughing.of.Blood    Fatigue Wheezing  Snoring
2          45.88091   5.693829 42.71864 176.6150
3          91.45923 104.380729 77.82862 139.1473

Std. Errors:
  (Intercept) Alcohol.use Dust.Allergy OccuPational.Hazards Genetic.Risk
2   415.05487    66.57656     70.36951            64.02858     82.63294
3    56.75419  1466.49193     32.60372           752.85603     32.60376
  Coughing.of.Blood   Fatigue Wheezing   Snoring
2          144.4591   89.8362 132.3439  64.05621
3         2398.6426 1876.6023 676.8353 444.17771
```

10.3 Appendix C

```
Model:
Level ~ Alcohol.use + Dust.Allergy + OccuPational.Hazards + Genetic.Risk +
    Coughing.of.Blood + Fatigue + Wheezing + Snoring
                     Df    AIC    LRT   Pr(Chi)
<none>                    36.02
Alcohol.use           2 118.74  86.71 < 2.2e-16 ***
Dust.Allergy          2 110.42  78.40 < 2.2e-16 ***
OccuPational.Hazards  2 156.28 124.26 < 2.2e-16 ***
Genetic.Risk          2 189.40 157.38 < 2.2e-16 ***
Coughing.of.Blood     2 145.09 113.07 < 2.2e-16 ***
Fatigue               2 375.80 343.78 < 2.2e-16 ***
Wheezing              2 186.22 154.20 < 2.2e-16 ***
Snoring               2 172.79 140.76 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

10.4 Appendix D

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	9.104	39.581	39.581	9.104	39.581	39.581	8.413	36.580	36.580
2	2.700	11.740	51.321	2.700	11.740	51.321	2.269	9.864	46.444
3	2.017	8.770	60.090	2.017	8.770	60.090	2.159	9.386	55.831
4	1.517	6.597	66.687	1.517	6.597	66.687	1.909	8.302	64.132
5	1.299	5.649	72.337	1.299	5.649	72.337	1.792	7.793	71.926
6	1.216	5.287	77.623	1.216	5.287	77.623	1.310	5.697	77.623
7	.791	3.437	81.060						
8	.691	3.003	84.064						
9	.605	2.632	86.695						
10	.509	2.215	88.910						
11	.499	2.171	91.082						
12	.418	1.817	92.899						
13	.340	1.479	94.377						
14	.282	1.226	95.603						
15	.226	.982	96.585						
16	.208	.903	97.488						
17	.156	.678	98.166						
18	.116	.504	98.669						
19	.096	.417	99.087						
20	.084	.367	99.453						
21	.069	.299	99.752						
22	.033	.145	99.896						
23	.024	.104	100.000						

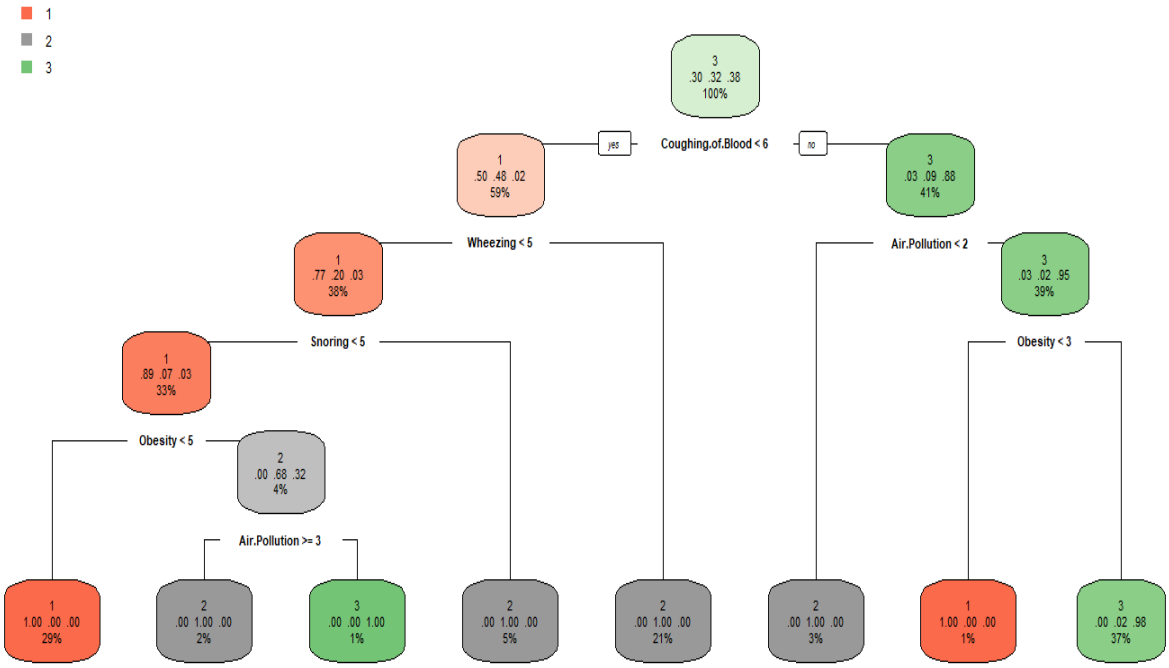
Extraction Method: Principal Component Analysis.

10.5 Appendix E

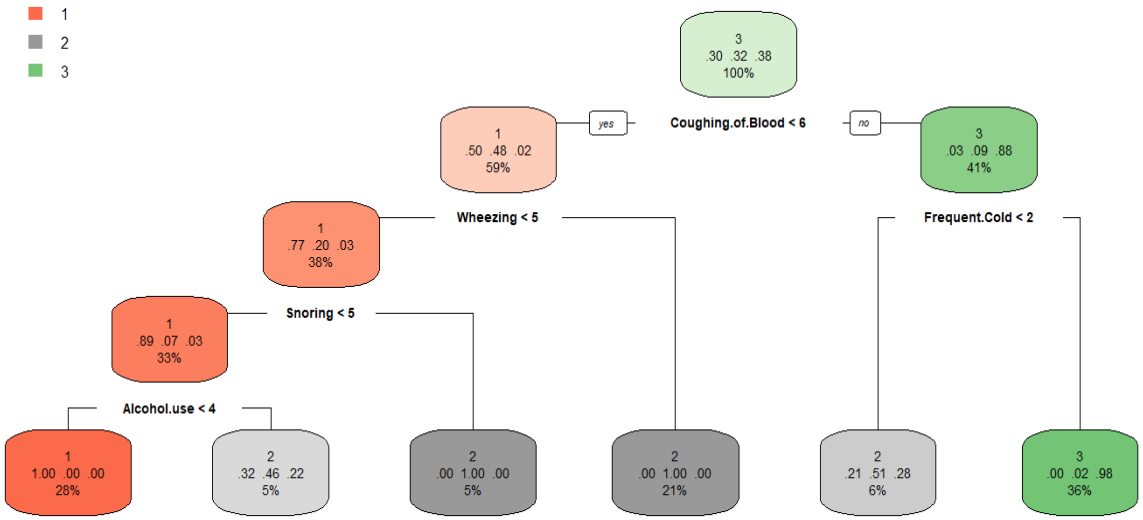
Rotated Component Matrix[a]

	Component					
	1	2	3	4	5	6
Age						.771
Gender						.719
AirPollution	.741					
Alcoholuse	.867					
DustAllergy	.773					
OccuPationalHazards	.890					
GeneticRisk	.902					
chronicLungDisease	.856					
BalancedDiet	.834					
Obesity	.793					
Smoking	.734					
PassiveSmoker	.780					
ChestPain	.902					
CoughingofBlood	.807					
Fatigue		.582	.424			
WeightLoss		.866				
ShortnessofBreath		.647				
Wheezing				.513	.690	
SwallowingDifficulty					.853	
ClubbingofFingerNails				.806		
FrequentCold			.831			
DryCough			.758			
Snoring		-.407	.578			

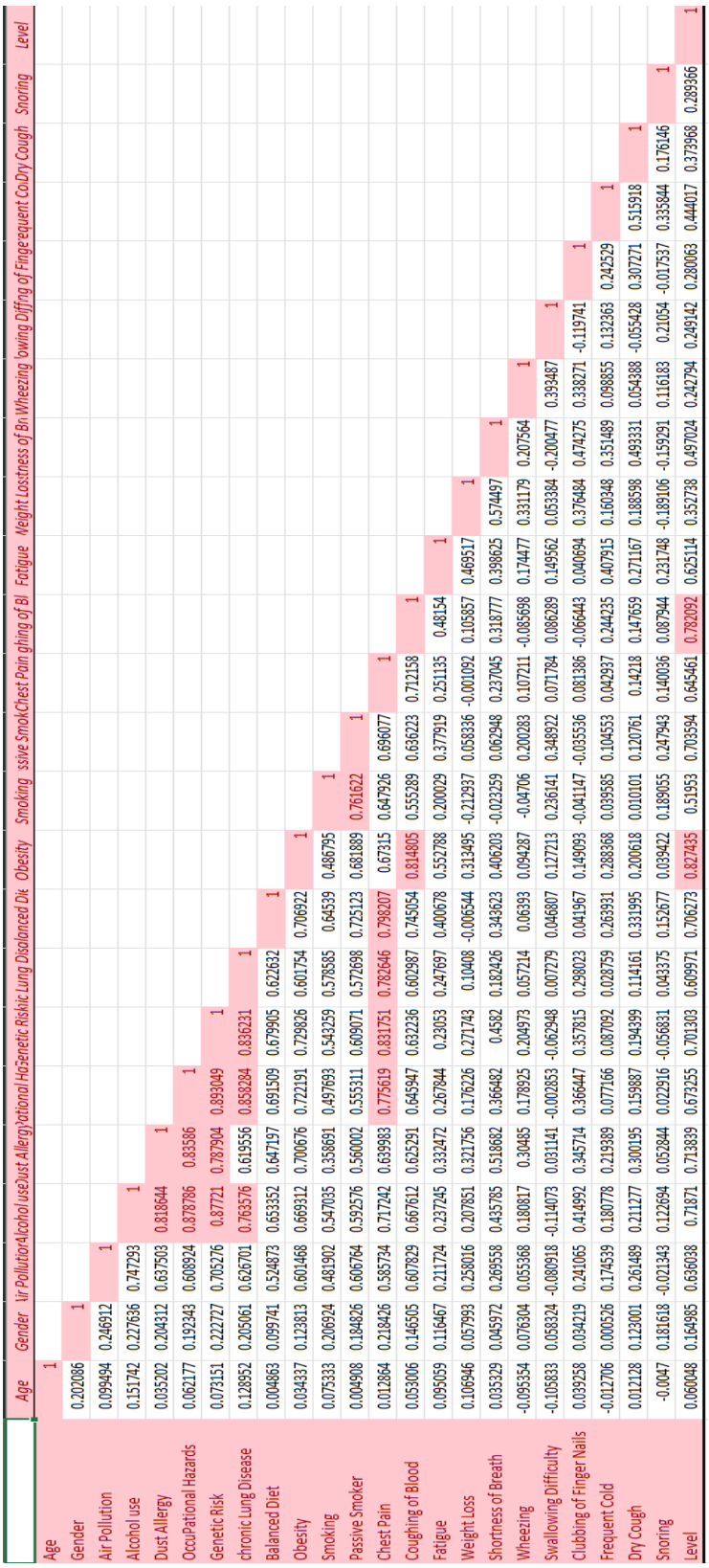
10.6 Appendix F



10.7 Appendix G



10.8 Appendix H



	Age	Gender	Air Pollution	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	Frequent Cold	Dry Cough	Snoring	Level
Age	1																							
Gender	0.202086	1																						
Air Pollution	0.099494	0.246912	1																					
Alcohol use	0.151742	0.227636	0.747293	1																				
Dust Allergy	0.035202	0.204312	0.637503	0.818644	1																			
OccuPational Hazards	0.062177	0.192343	0.608924	0.878786	0.83586	1																		
Genetic Risk	0.073151	0.222727	0.705276	0.87721	0.787904	0.893049	1																	
chronic Lung Disease	0.128952	0.205061	0.626701	0.763576	0.619556	0.858284	0.836231	1																
Balanced Diet	0.004863	0.099741	0.524873	0.653352	0.647197	0.691509	0.679905	0.622632	1															
Obesity	0.034337	0.123813	0.601468	0.669312	0.700676	0.722191	0.729826	0.601754	0.706922	1														
Smoking	0.075333	0.206924	0.481902	0.547035	0.358691	0.497693	0.543259	0.578585	0.64539	0.486795	1													
Passive Smoker	0.004908	0.184826	0.606764	0.592576	0.560002	0.555311	0.609071	0.572698	0.725123	0.681889	0.761622	1												
Chest Pain	0.012864	0.218426	0.585734	0.717242	0.639983	0.775619	0.831751	0.782646	0.798207	0.67315	0.647926	0.696077	1											
Coughing of Blood	0.053006	0.146505	0.607829	0.667612	0.625291	0.645947	0.632236	0.602987	0.745054	0.814805	0.555289	0.636223	0.712158	1										
Fatigue	0.095059	0.116467	0.211724	0.237245	0.332472	0.267844	0.23053	0.247697	0.400678	0.552788	0.200029	0.377919	0.251135	0.48154	1									
Weight Loss	0.106946	0.057993	0.258016	0.207851	0.321756	0.176226	0.271743	0.10408	-0.006544	0.313495	-0.212937	0.058336	-0.001092	0.105857	0.469517	1								
Shortness of Breath	0.035329	0.045972	0.269558	0.435785	0.518682	0.366482	0.4582	0.182426	0.343623	0.406203	-0.023259	0.062948	0.237045	0.318777	0.398625	0.574497	1							
Wheezing	-0.095354	0.076304	0.055368	0.180817	0.30485	0.178925	0.204973	0.057214	0.06393	0.094287	-0.04706	0.200283	0.107211	-0.085698	0.174477	0.331179	0.207564	1						
Swallowing Difficulty	-0.105833	0.058324	-0.080918	-0.114073	0.031141	-0.002853	-0.062948	0.007279	0.046807	0.127213	0.226141	0.248922	0.071784	0.086289	0.149562	0.053384	-0.200477	0.393487	1					
Clubbing of Finger Nails	0.039258	0.034219	0.241065	0.414992	0.345714	0.366447	0.357815	0.298023	0.041967	0.149093	-0.041147	-0.035536	0.081386	-0.066443	0.040694	0.376484	0.474275	0.338271	-0.119741	1				
Frequent Cold	-0.012706	0.000526	0.174539	0.180778	0.219389	0.077166	0.087092	0.028759	0.263931	0.288368	0.039585	0.104533	0.042937	0.244235	0.407915	0.160348	0.351489	0.098855	0.132363	0.242529	1			
Dry Cough	0.012128	0.123001	0.261489	0.211277	0.300195	0.159887	0.194399	0.114161	0.331995	0.200618	0.010101	0.120761	0.14218	0.147659	0.271167	0.188598	0.493331	0.054388	-0.055428	0.307271	0.515918	1		
Snoring	-0.0047	0.181618	-0.021343	0.122694	0.052844	0.022916	-0.056831	0.043375	0.152677	0.039422	0.189055	0.247943	0.140036	0.087944	0.231748	-0.189106	-0.159291	0.116183	0.21054	-0.017537	0.335844	0.176146	1	
Level	0.060048	0.164985	0.636038	0.71871	0.713839	0.673255	0.701303	0.609971	0.706273	0.827435	0.51953	0.703594	0.645461	0.782092	0.625114	0.352738	0.497024	0.242794	0.249142	0.280063	0.444017	0.373968	0.289366	1

10.9 Appendix I: Meme