PDANA8411 POE PART 2

Brice Derek Agnew (ST10072411)

IIE VARSITY COLLEGE Cape Town

Contents

Dataset and Models:

The dataset used was the "**Lung Cancer Prediction**"(The Devastator, 2021) Dataset, provided by the user "**The Devastator**" on Kaggle.com, available using the following link [Lung Cancer Prediction] I found this dataset suitable for our cancer classification algorithm since it included a clear categorical target variable in the form of 'Level', more specifically 'level of cancer risk', it is a supervised learning dataset, and it includes mixed datatypes. Additionally, after evaluating similar datasets of the Kaggle platform, it was found that similar datasets dealt in determining the severity of existing cancers, whereas I thought that this project would be far more useful in real world applications as a 'cancer screening' program, that determines if a person has cancer at all. I thought that this was important as the 'severity' focussed datasets posed the risk of over diagnosing someone with a cancer that they might not have.

The models I decided on using were the 'Random Forest' classification model and the 'XGBoost' classification model.

I decided that the Random Forest model was suitable for this type of dataset, since the relationships were not guaranteed to be linear, the datatypes were mixed, the dimensionality of the dataset was between 16 and 30, and this was an example of a classification problem.

I also chose to use XGBoost as the second model I would train, since it had similar benefits to the random forest model, in terms of the non-linear adaptability, mixed data type compatibility and acceptable dimensionality range, but was slightly more sensitive to hyper parameters and was slightly slower to train.

Plan:

The data analysis that will be done for this report will be structured in the following steps:

- Data Collection: the data used will be collected from a relevant selected set from Kaggle.com, namely the "**Lung Cancer Prediction**" Dataset , due to its relevance to the topic of study and relevant target value of 'Level'.

- Data Cleaning: The dataset will be investigated to ensure that there are no duplicate or missing values

    1. Any missing values found will have their entries deleted.
    2. Any duplicate values found will have their entries deleted.
    3. Categorical values will be encoded to interface with the Machine learning models.

- Exploratory Data Analysis: Key statistics and distributions will be identified and presented using visual representations.
    1. Initial key statistics such as entry count, standard deviation, mean, upper and lower bounds and maximum observed value will be identified and displayed for every feature in the set.
    2. The distributions of each feature will be identified and displayed using histogram graphs.
    3. The correlation between features will be identified and represented visually using a heatmap plot.

- Feature Selection and Data Splitting: After the exploratory data analysis is complete, the features will then be broken into training and test sets respectively, and will be subjected to various tests in order to determine the relevance of each feature to inform dimensionality reduction. The relevance metrics calculated will be the Pearson coefficient, the Chi squared value, and the p-value of each feature. After these values are calculated, the training set will be trimmed in order to remove the least relevant features from the training and test data, based on chi scores.

- Model Training: After feature selection is completed, the models will then be trained utilizing pipeline architecture and the adjusted training data.

- Model Evaluation and Comparison: The accuracy of each model will be evaluated based on several key statistics and formats. The confusion matrix, Cross validation, and a classification report table, consisting of precision, recall, F1, and support scores, as well as accuracy, averages and weighted averages.

Report:

The Data cleaning process found that there were zero missing values and no further action was needed.

```
Missing Values per column:
Patient Id                  0
Age                         0
Gender                      0
Air Pollution               0
Alcohol use                 0
Dust Allergy                0
OccuPational Hazards        0
Genetic Risk                0
chronic Lung Disease        0
Balanced Diet               0
Obesity                     0
Smoking                     0
Passive Smoker              0
Chest Pain                  0
Coughing of Blood           0
Fatigue                     0
Weight Loss                 0
Shortness of Breath         0
Wheezing                    0
Swallowing Difficulty       0
Clubbing of Finger Nails    0
Frequent Cold               0
Dry Cough                   0
Snoring                     0
Level                       0
dtype: int64

 Total number of missing values in set: 0
```

It was found that there were no duplicate columns or rows, and no further action was needed.

```
Total number of Duplicated rows: 0


Total number of Duplicate columns: 0
```

The columns of the dataset were renamed to be in lowercase and no spaces for consistency.

	age	gender	air_pollution	alcohol_use	dust_allergy	occupational_hazards	genetic_risk	chronic_lung_disease	balanced_diet	obesity	...
count	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	...
mean	37.174000	1.402000	3.8400	4.563000	5.165000	4.840000	4.580000	4.380000	4.491000	4.465000	...
std	12.005493	0.490547	2.0304	2.620477	1.980833	2.107805	2.126999	1.848518	2.135528	2.124921	...
min	14.000000	1.000000	1.0000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...
25%	27.750000	1.000000	2.0000	2.000000	4.000000	3.000000	2.000000	3.000000	2.000000	3.000000	...
50%	36.000000	1.000000	3.0000	5.000000	6.000000	5.000000	5.000000	4.000000	4.000000	4.000000	...
75%	45.000000	2.000000	6.0000	7.000000	7.000000	7.000000	7.000000	6.000000	7.000000	7.000000	...
max	73.000000	2.000000	8.0000	8.000000	8.000000	8.000000	7.000000	7.000000	7.000000	7.000000	...

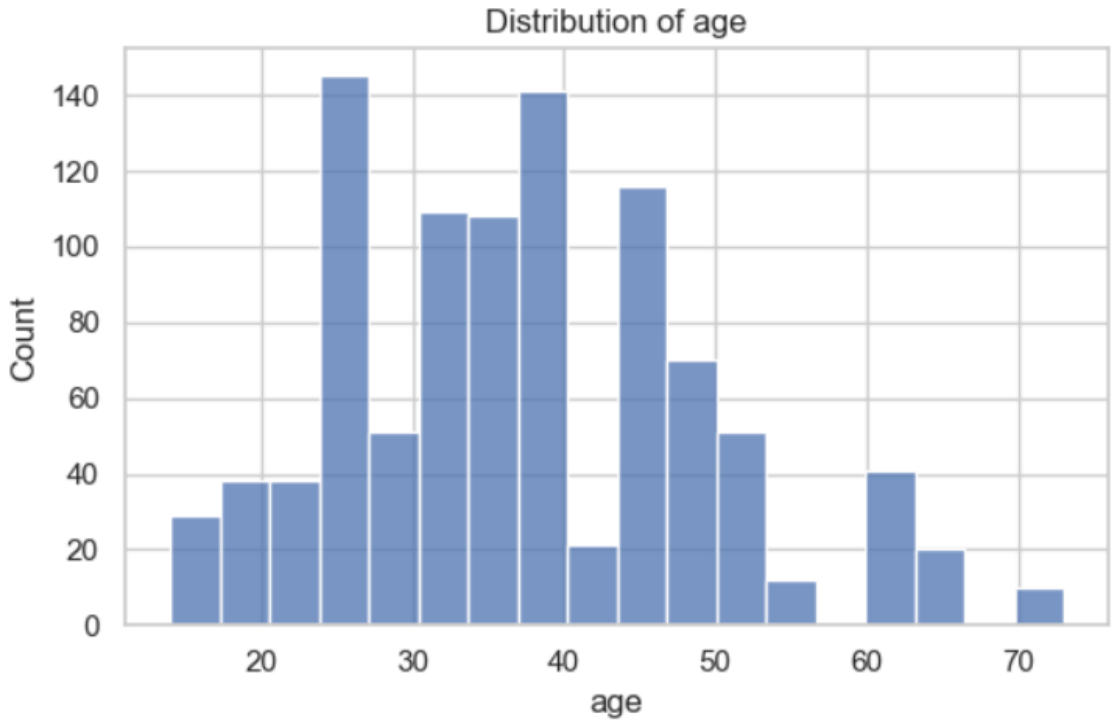
8 rows × 23 columns

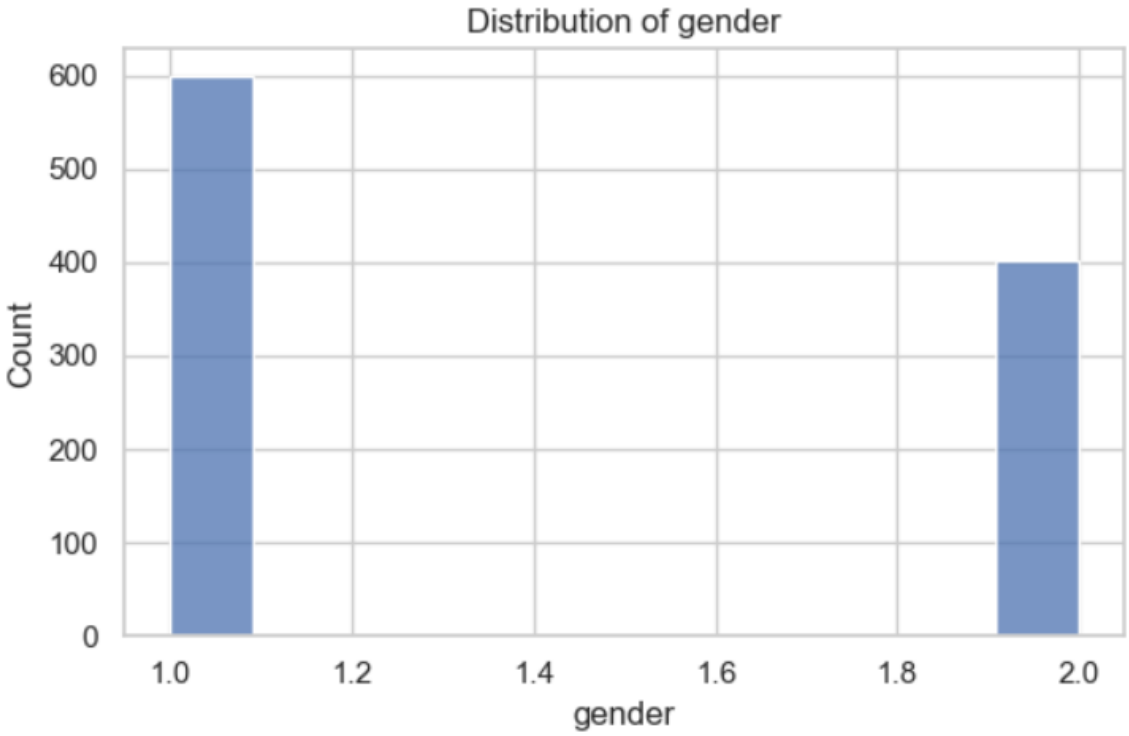
Key statistics for each feature were identified.

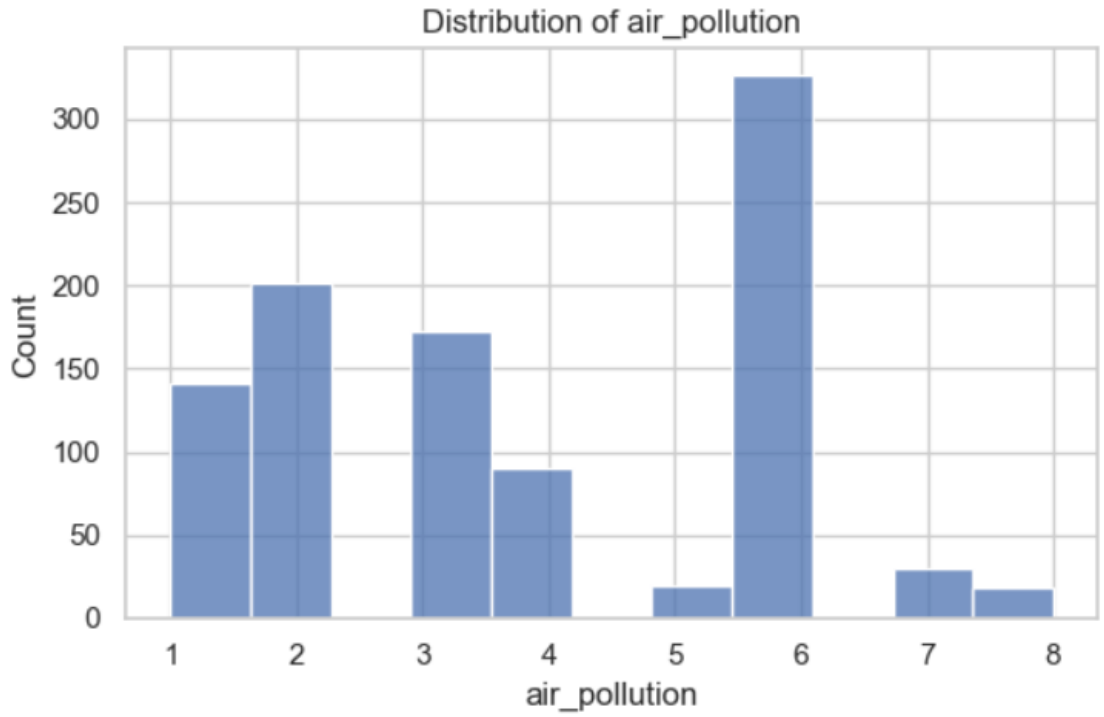
	age	gender	air_pollution	alcohol_use	dust_allergy	occupational_hazards	genetic_risk	chronic_lung_disease	balanced_diet	obesity	...
count	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	...
mean	37.174000	1.402000	3.8400	4.563000	5.165000	4.840000	4.580000	4.380000	4.491000	4.465000	...
std	12.005493	0.490547	2.0304	2.620477	1.980833	2.107805	2.126999	1.848518	2.135528	2.124921	...
min	14.000000	1.000000	1.0000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...
25%	27.750000	1.000000	2.0000	2.000000	4.000000	3.000000	2.000000	3.000000	2.000000	3.000000	...
50%	36.000000	1.000000	3.0000	5.000000	6.000000	5.000000	5.000000	4.000000	4.000000	4.000000	...
75%	45.000000	2.000000	6.0000	7.000000	7.000000	7.000000	7.000000	6.000000	7.000000	7.000000	...
max	73.000000	2.000000	8.0000	8.000000	8.000000	8.000000	7.000000	7.000000	7.000000	7.000000	...

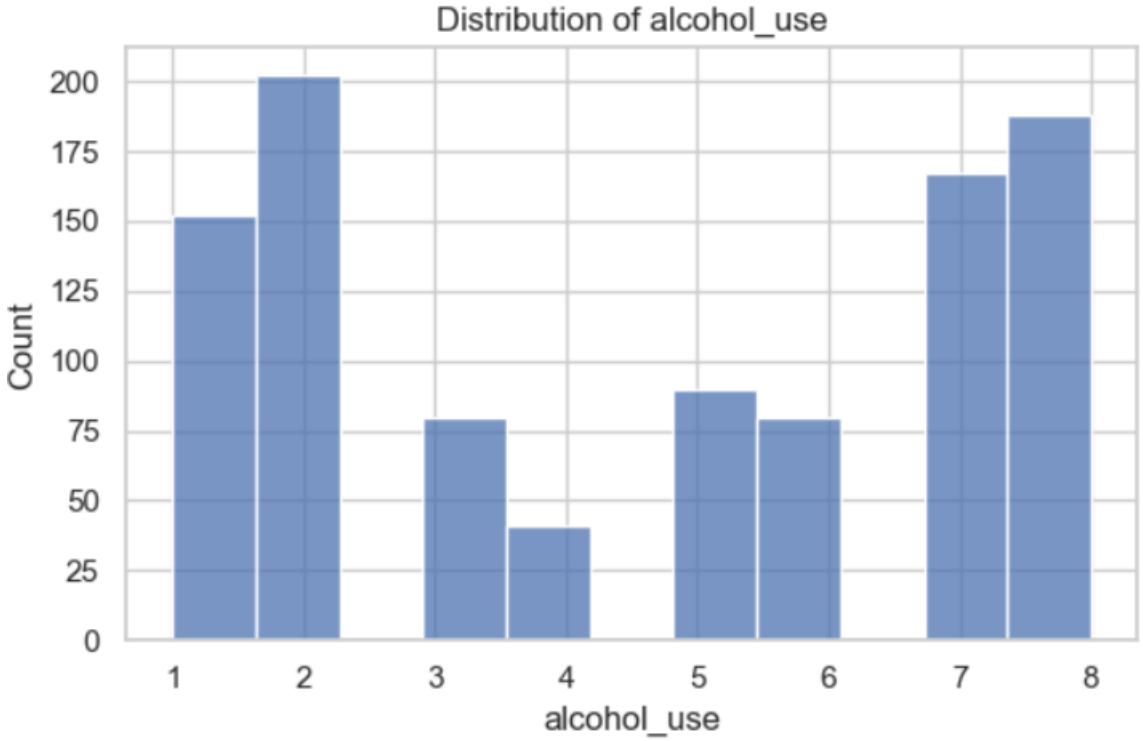
8 rows × 23 columns

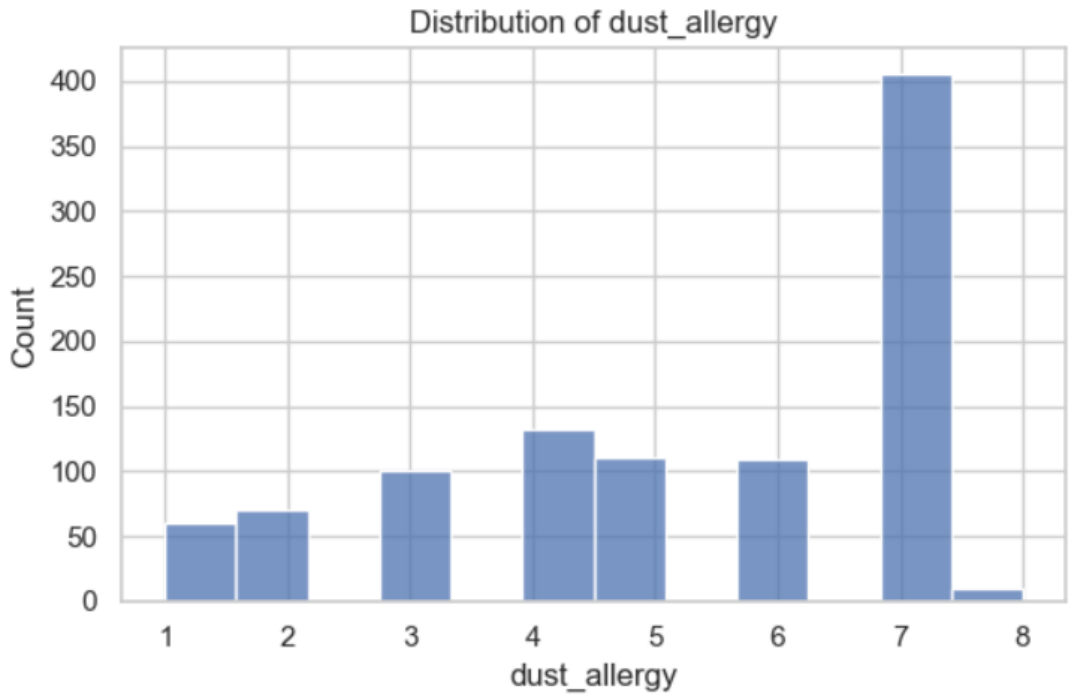
The data visualizations for each of the features are as follows:

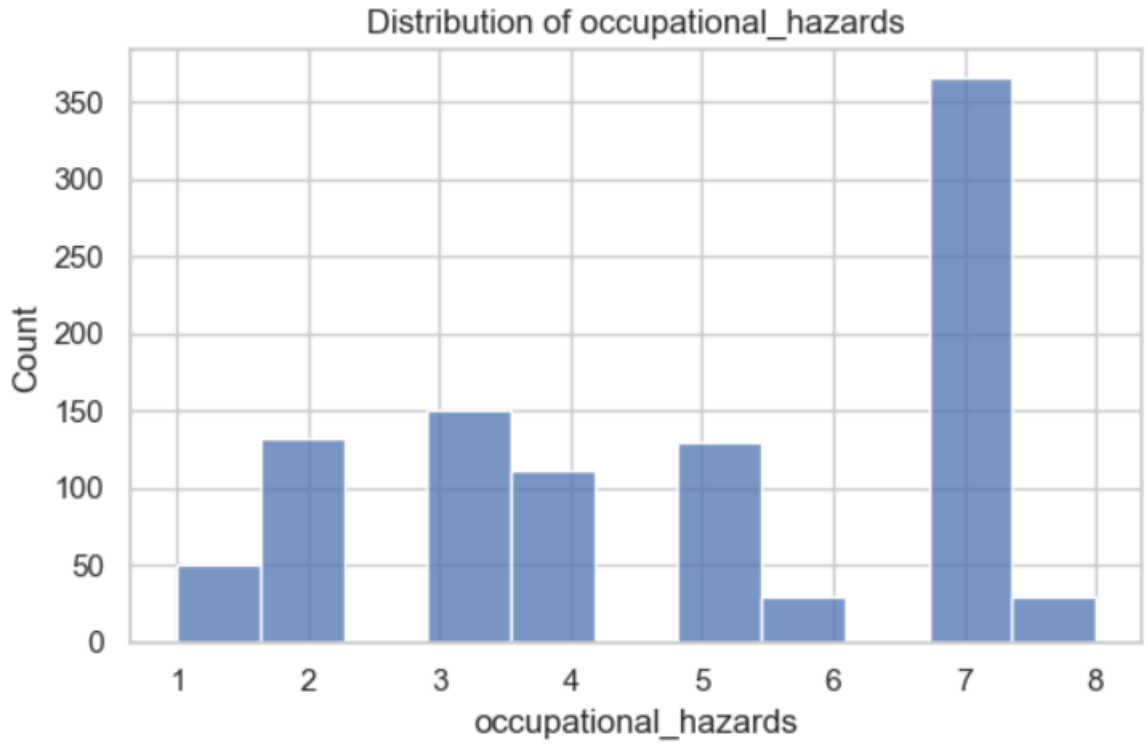


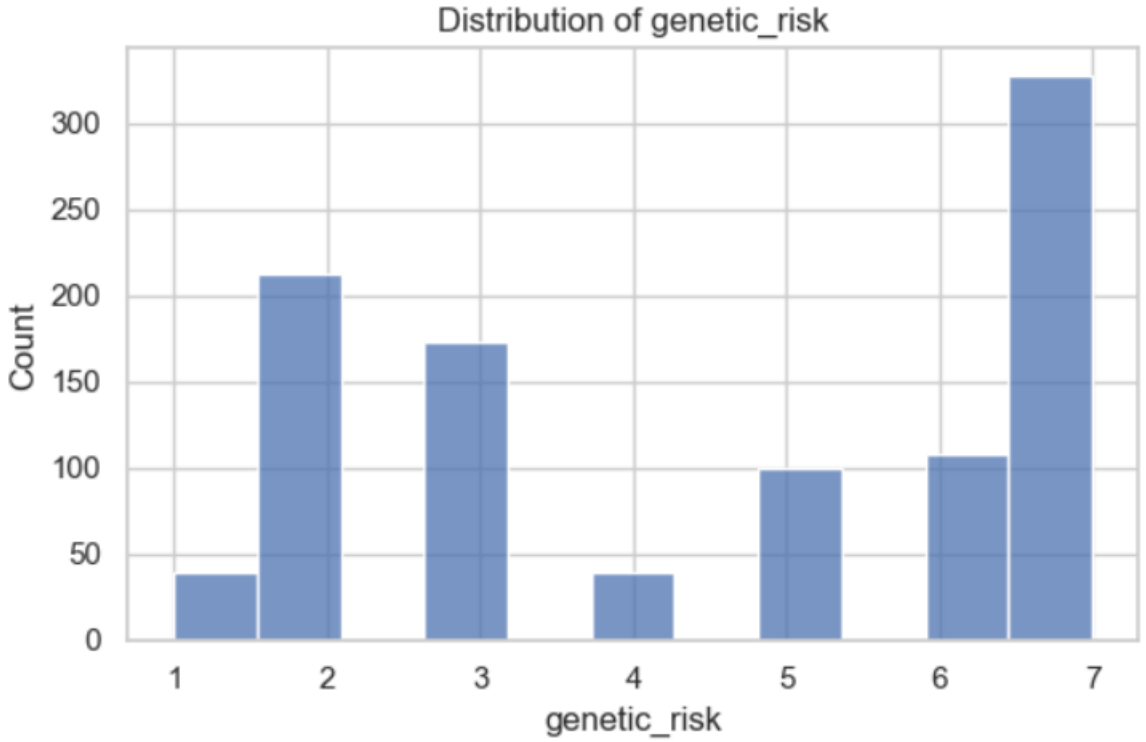


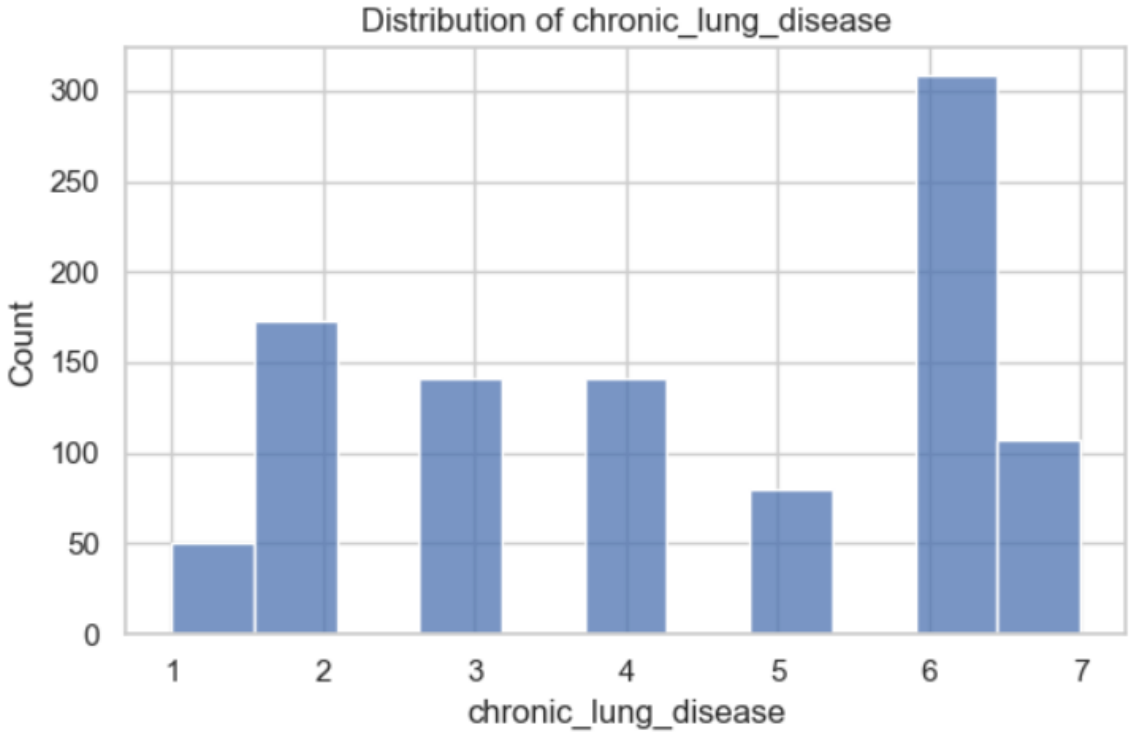

Distribution of air_pollution
Count
0
50
100
150
200
250
300
1
2
3
4
5
6
7
8
air_pollution

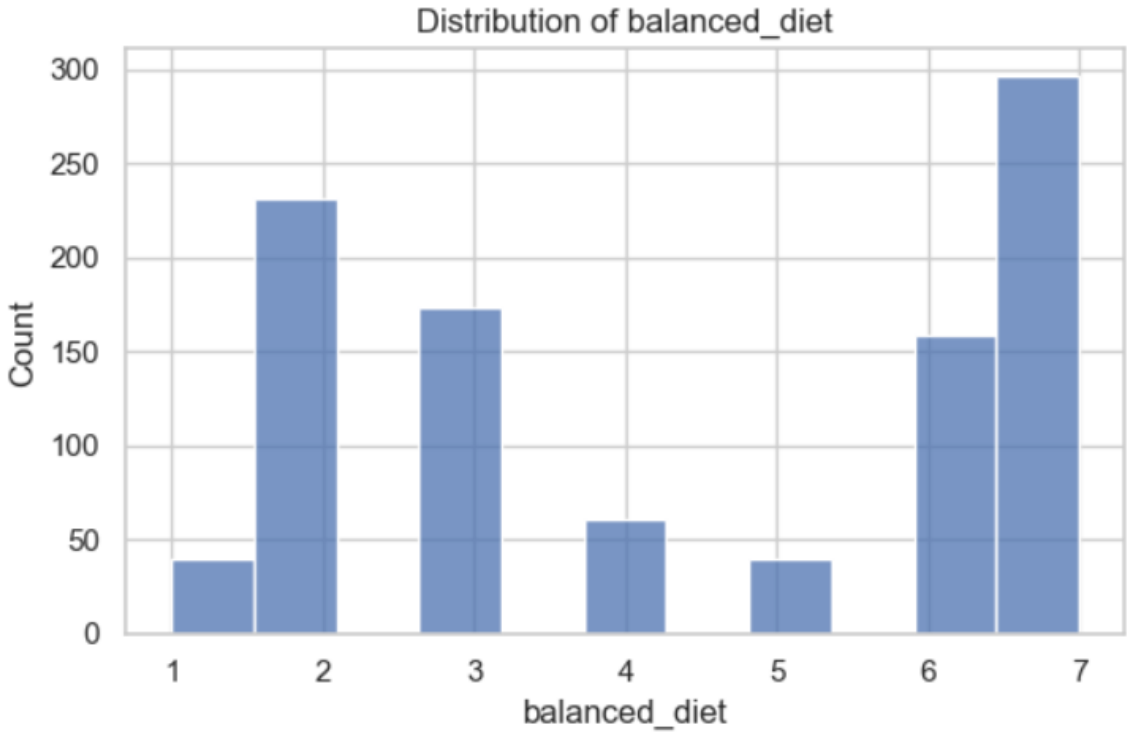

Distribution of alcohol_use
Count
0
25
50
75
100
125
150
175
200
1
2
3
4
5
6
7
8
alcohol_use

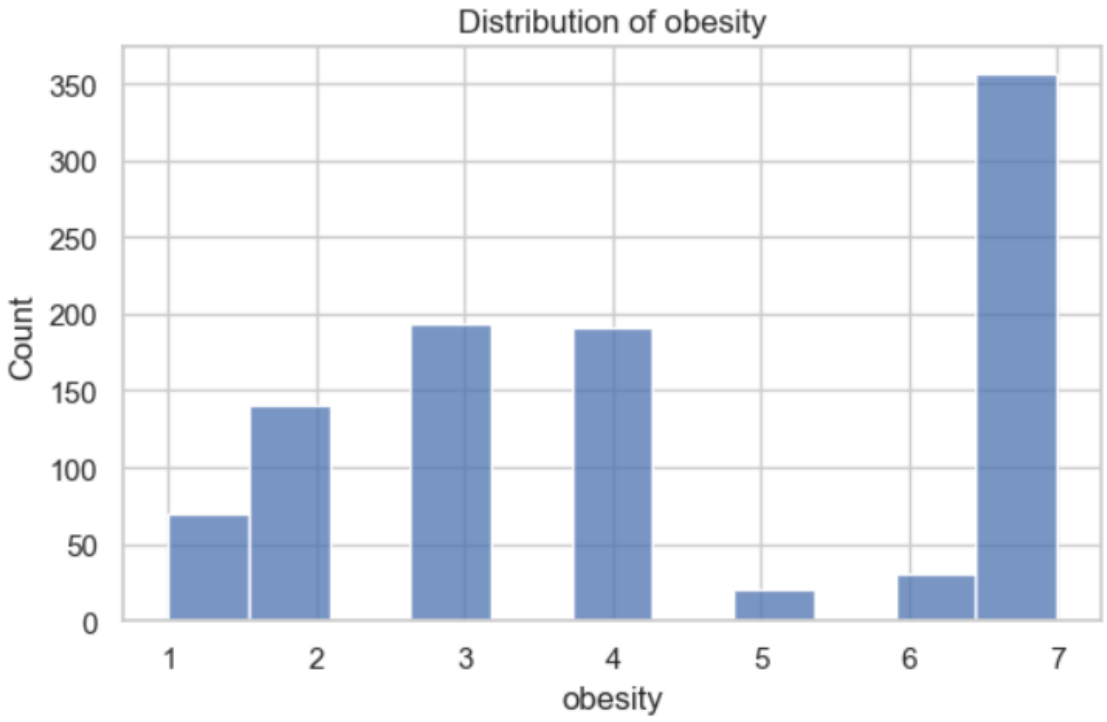

Distribution of dust_allergy
Count
400
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
dust_allergy

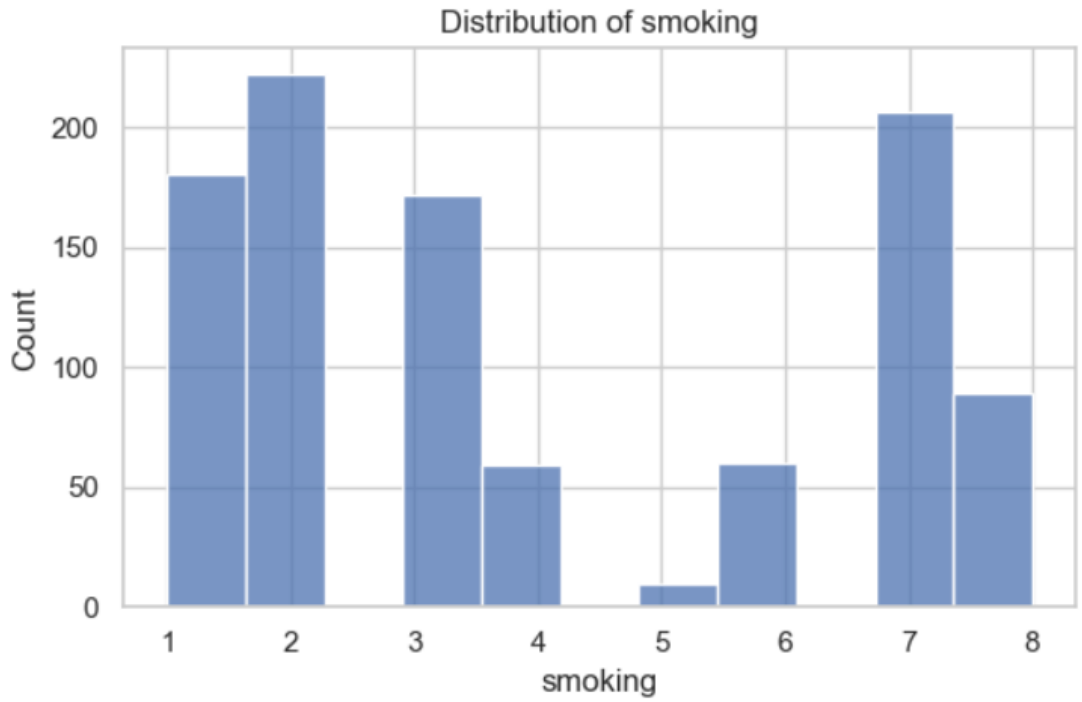

Distribution of occupational_hazards
Count
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
occupational_hazards

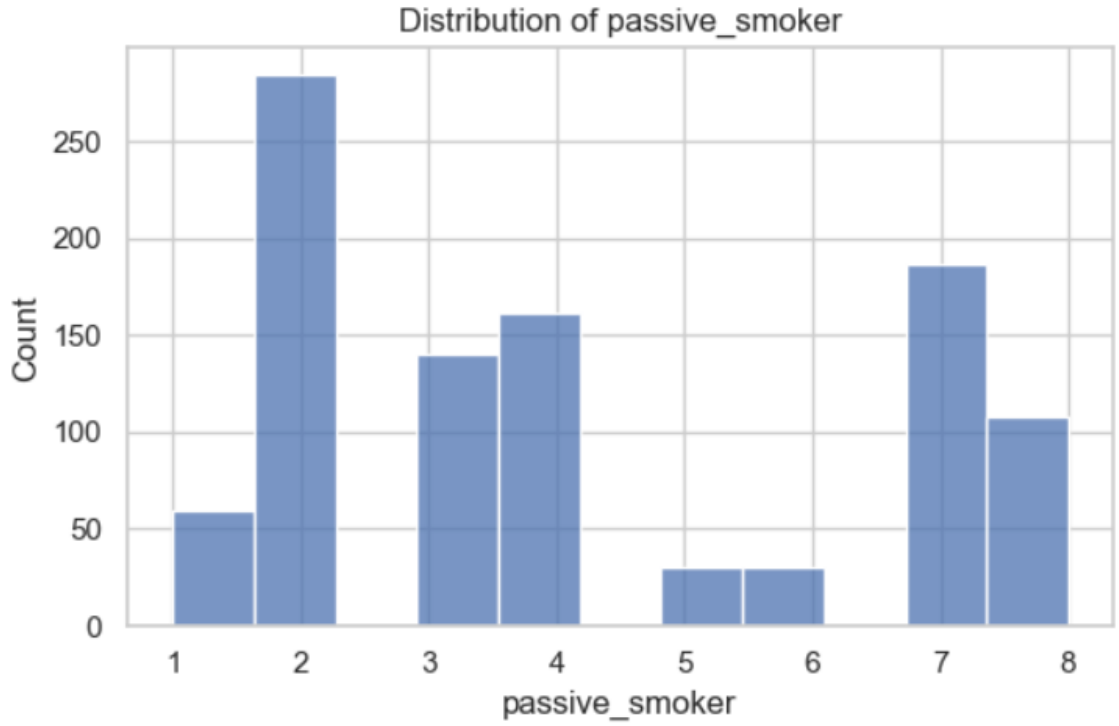

Distribution of genetic_risk
Count
0
50
100
150
200
250
300
1
2
3
4
5
6
7
genetic_risk

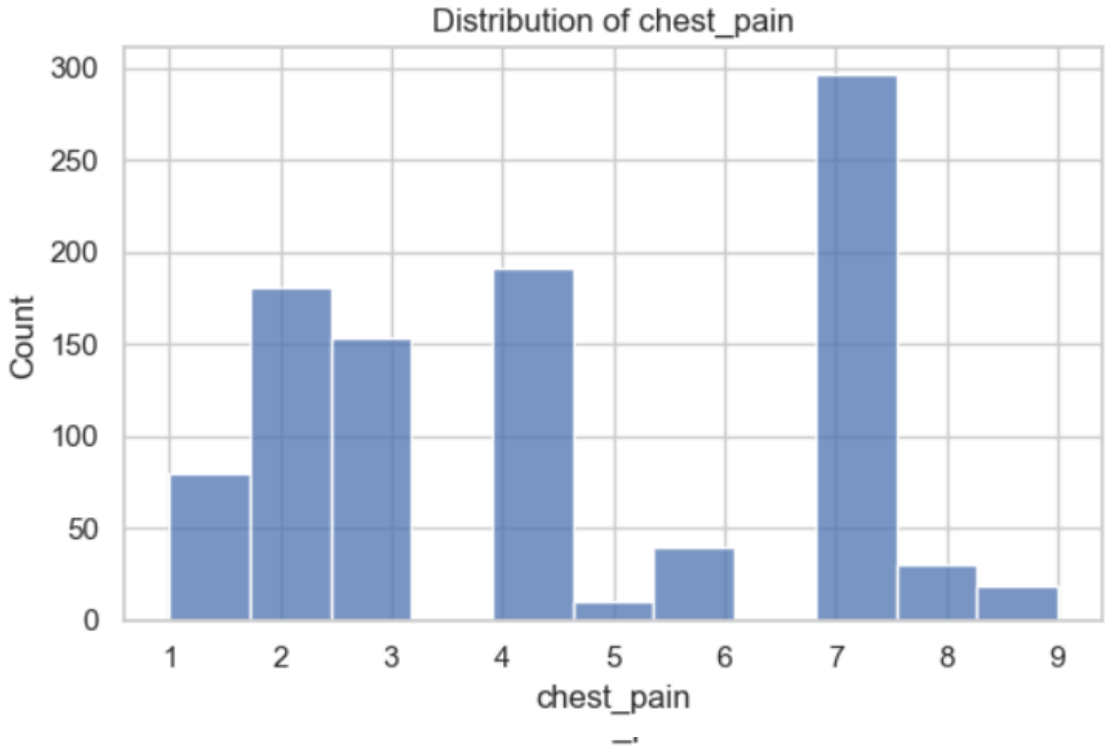

Distribution of chronic_lung_disease
Count
0
50
100
150
200
250
300
1
2
3
4
5
6
7
chronic_lung_disease

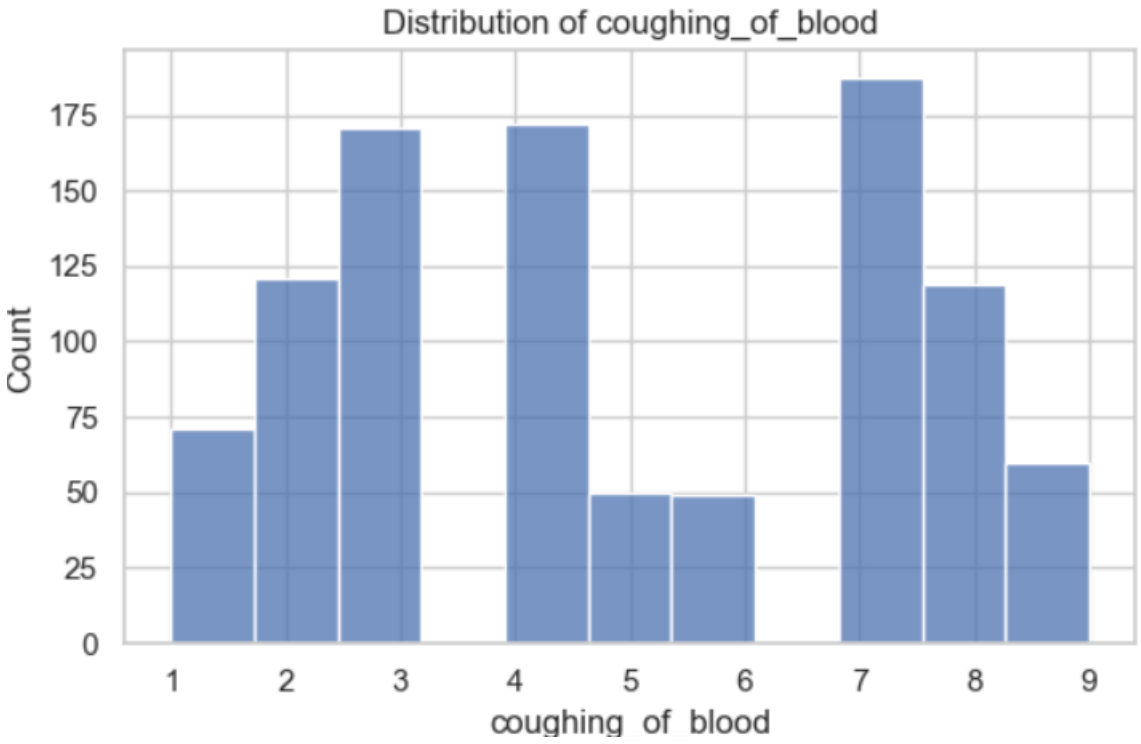

Distribution of balanced_diet
Count
300
250
200
150
100
50
0
1
2
3
4
5
6
7
balanced_diet

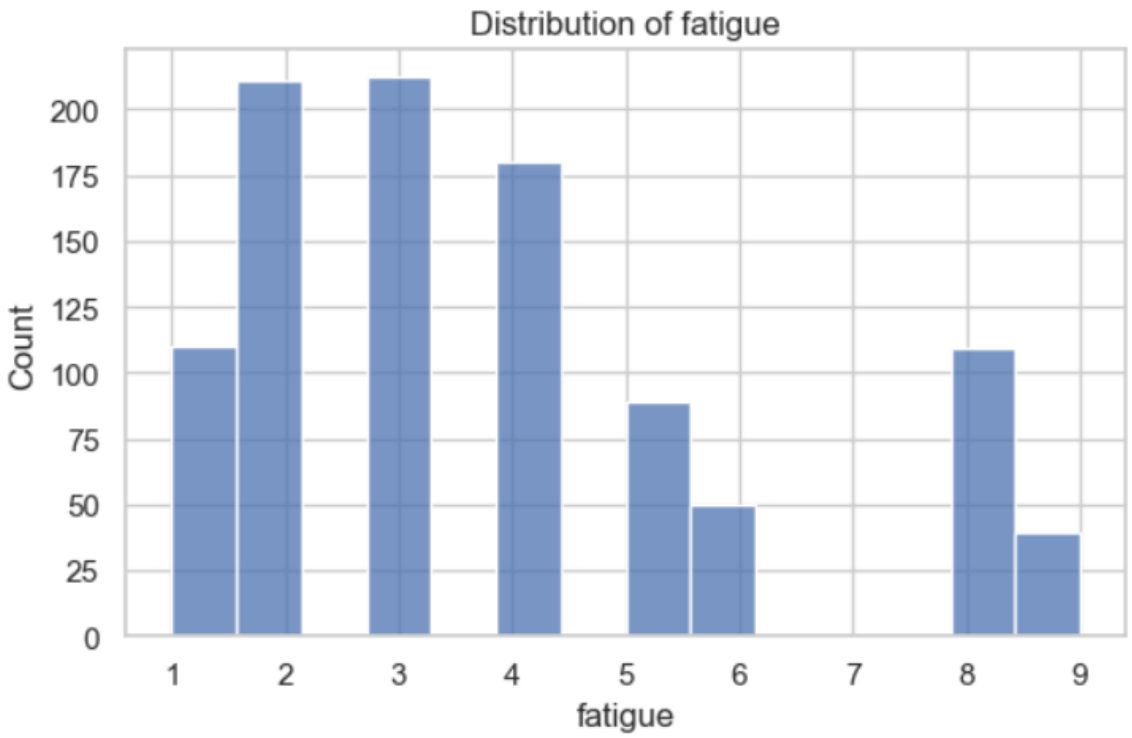

Distribution of obesity
Count
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
obesity

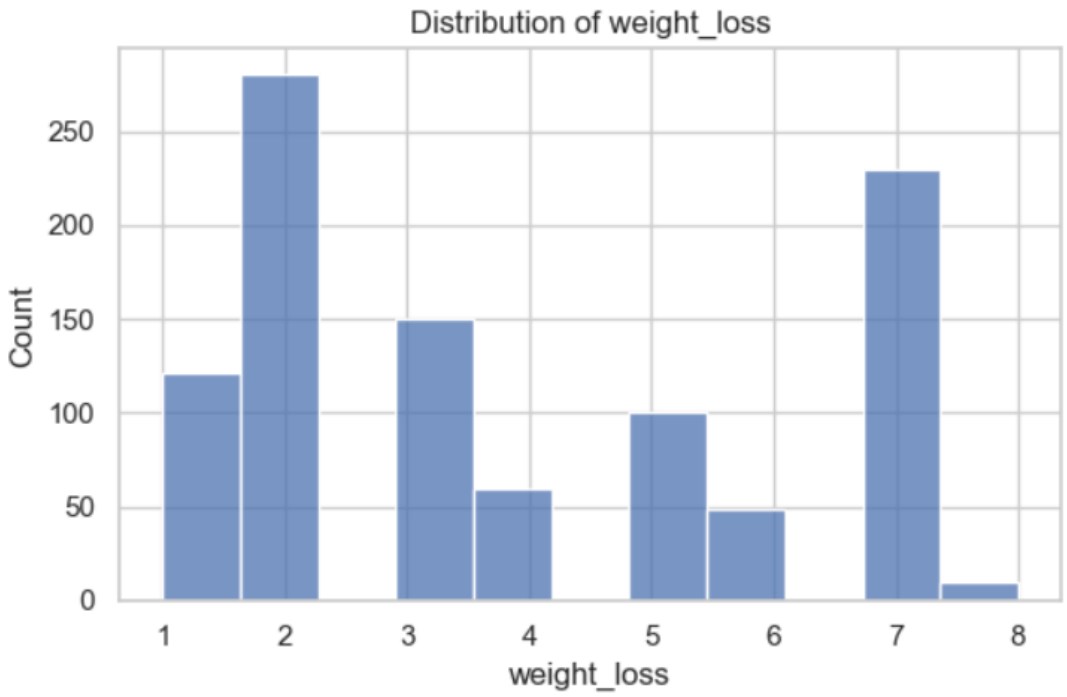

Distribution of smoking
Count
0
50
100
150
200
1
2
3
4
5
6
7
8
smoking

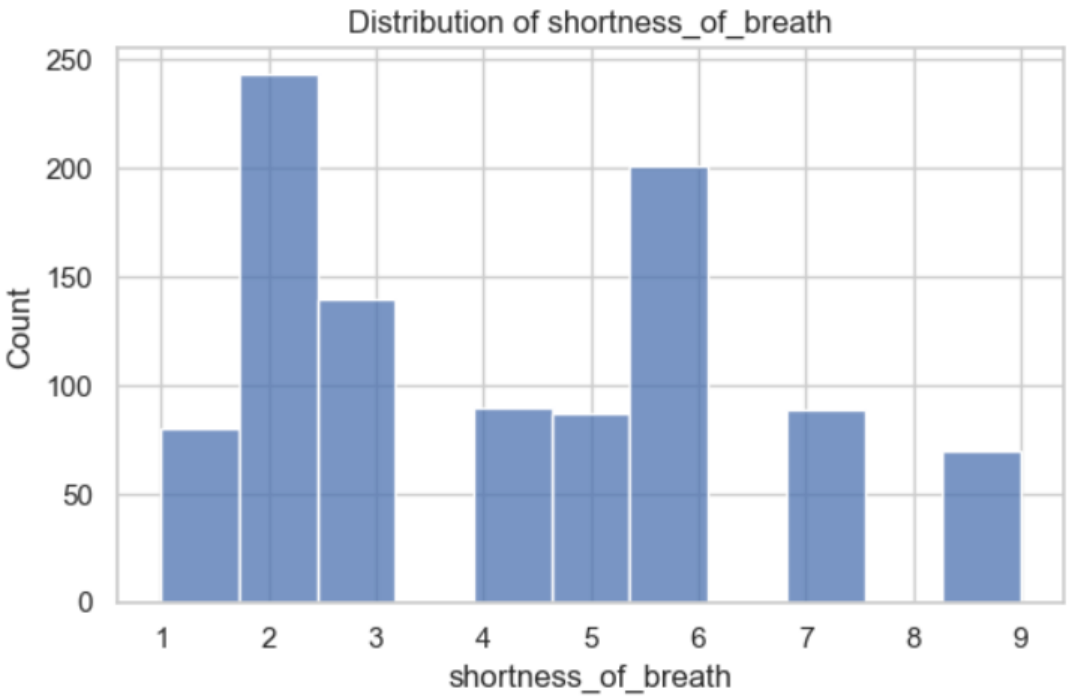

Distribution of passive_smoker
Count
0
50
100
150
200
250
1
2
3
4
5
6
7
8
passive_smoker

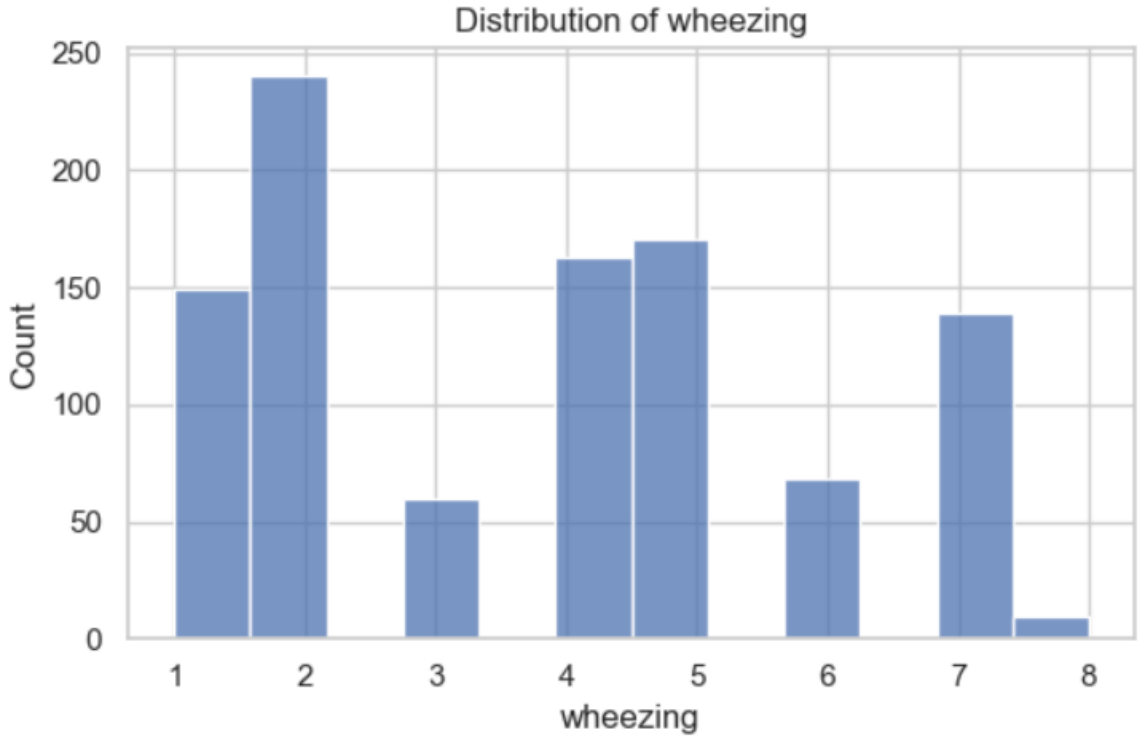

Distribution of chest_pain
Count
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
chest_pain

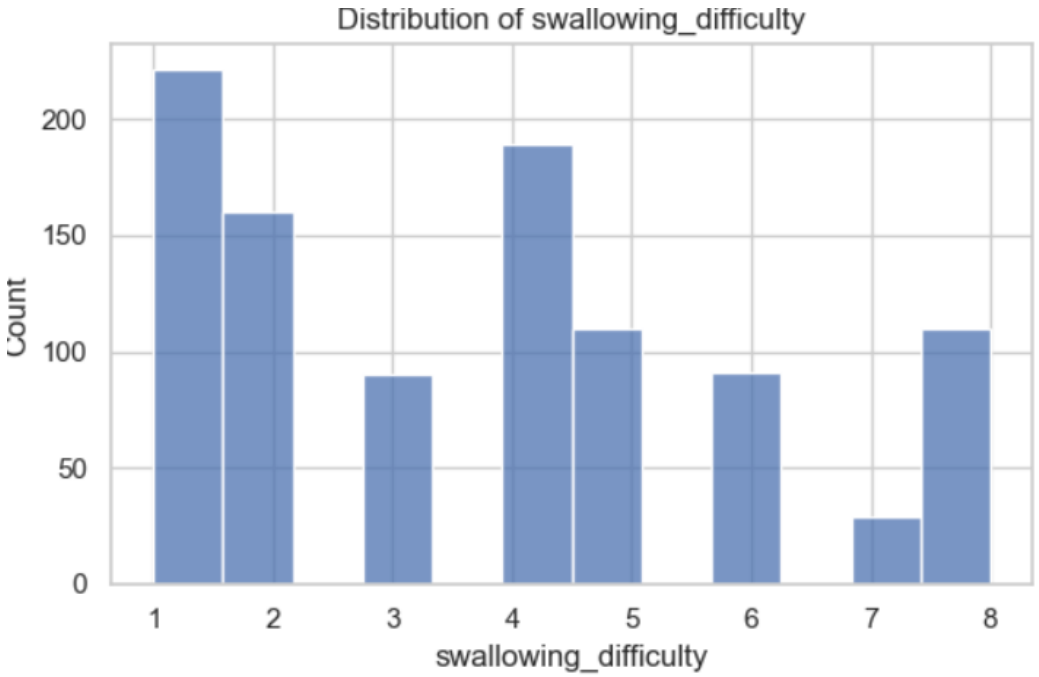

Distribution of coughing_of_blood
Count
175
150
125
100
75
50
25
0
1
2
3
4
5
6
7
8
9
coughing_of_blood

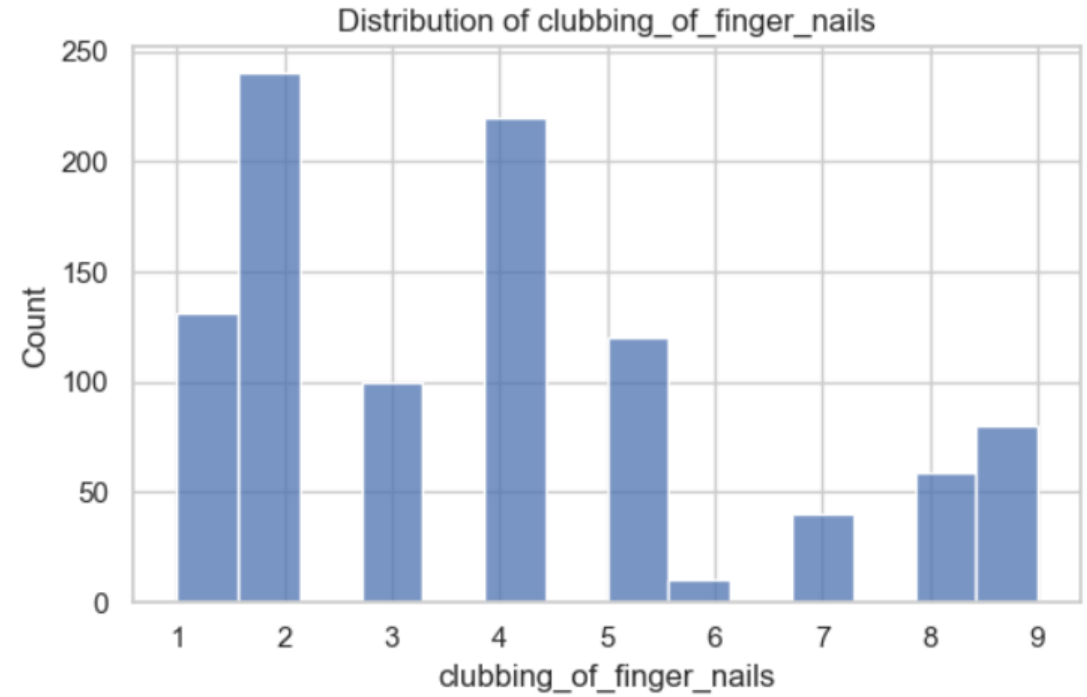

Distribution of fatigue
Count
200
175
150
125
100
75
50
25
0
1
2
3
4
5
6
7
8
9
fatigue

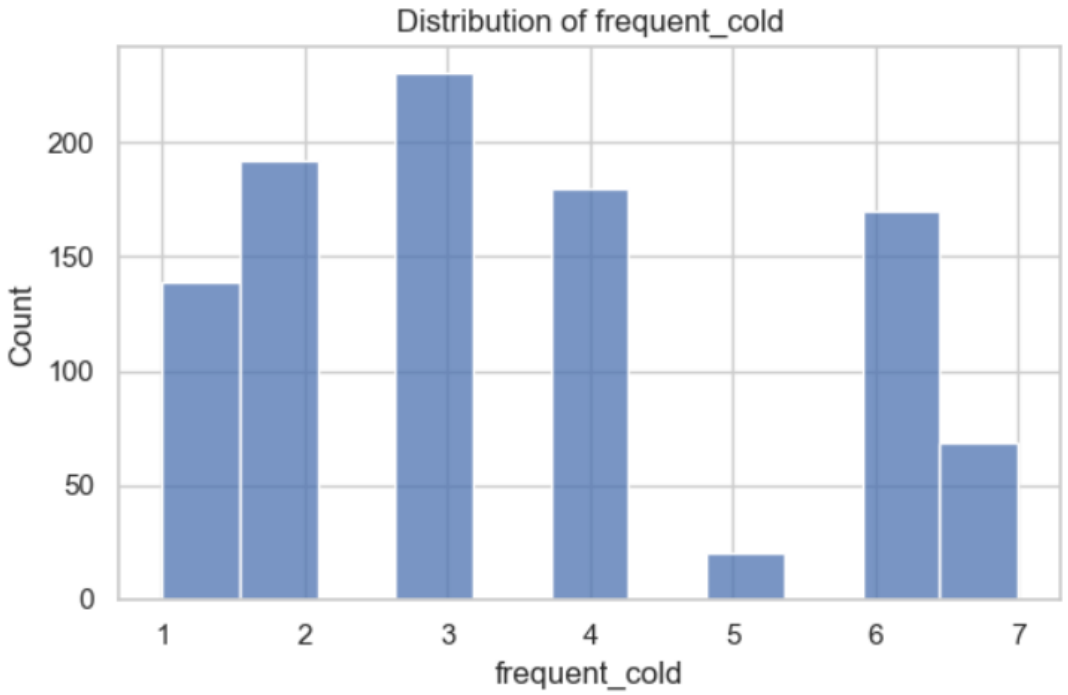

Distribution of weight_loss
Count
250
200
150
100
50
0
1
2
3
4
5
6
7
8
weight_loss

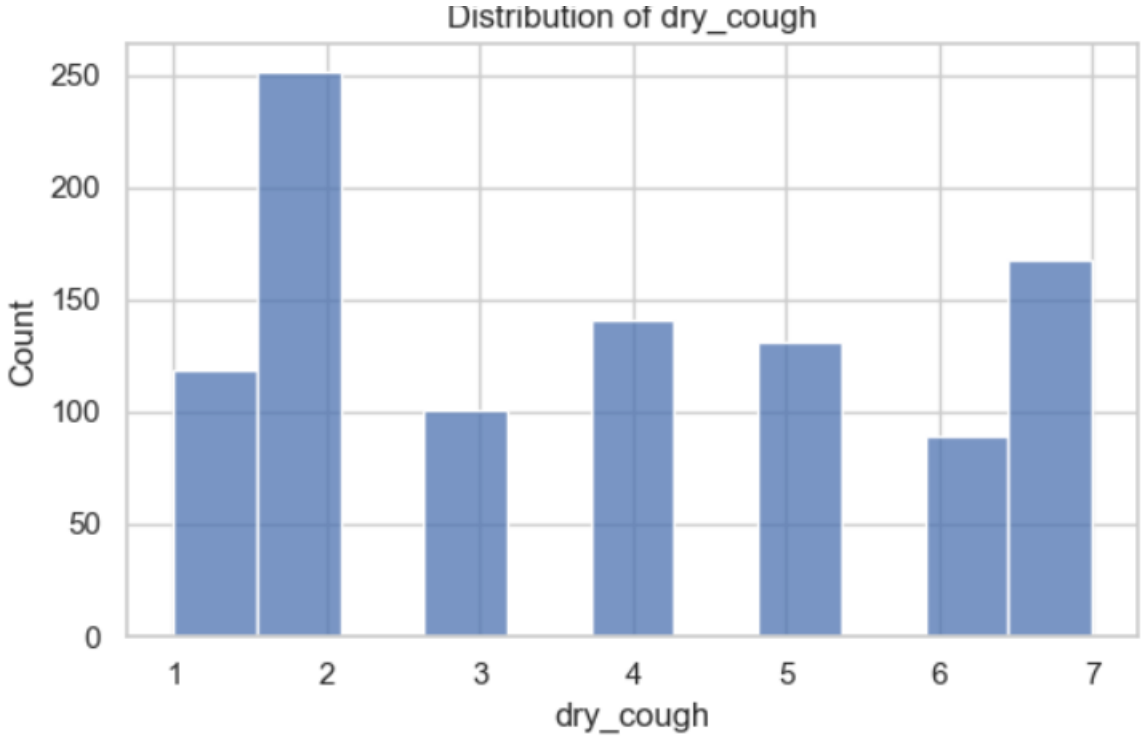

Distribution of shortness_of_breath
Count
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
shortness_of_breath

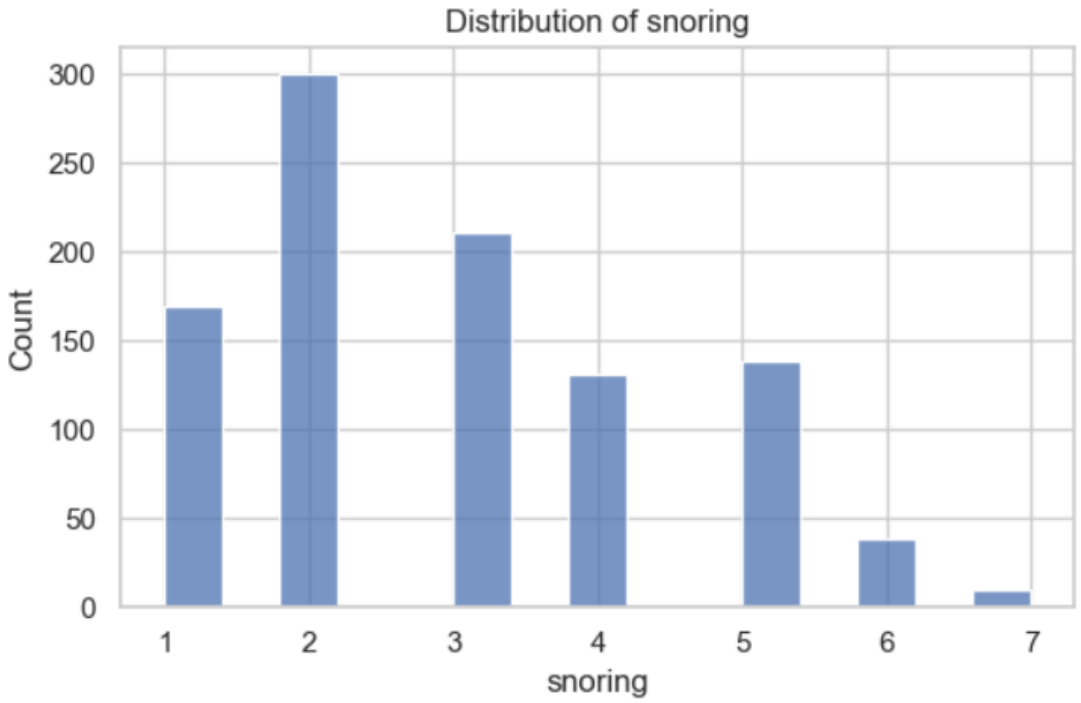

Distribution of wheezing
Count
250
200
150
100
50
0
1
2
3
4
5
6
7
8
wheezing

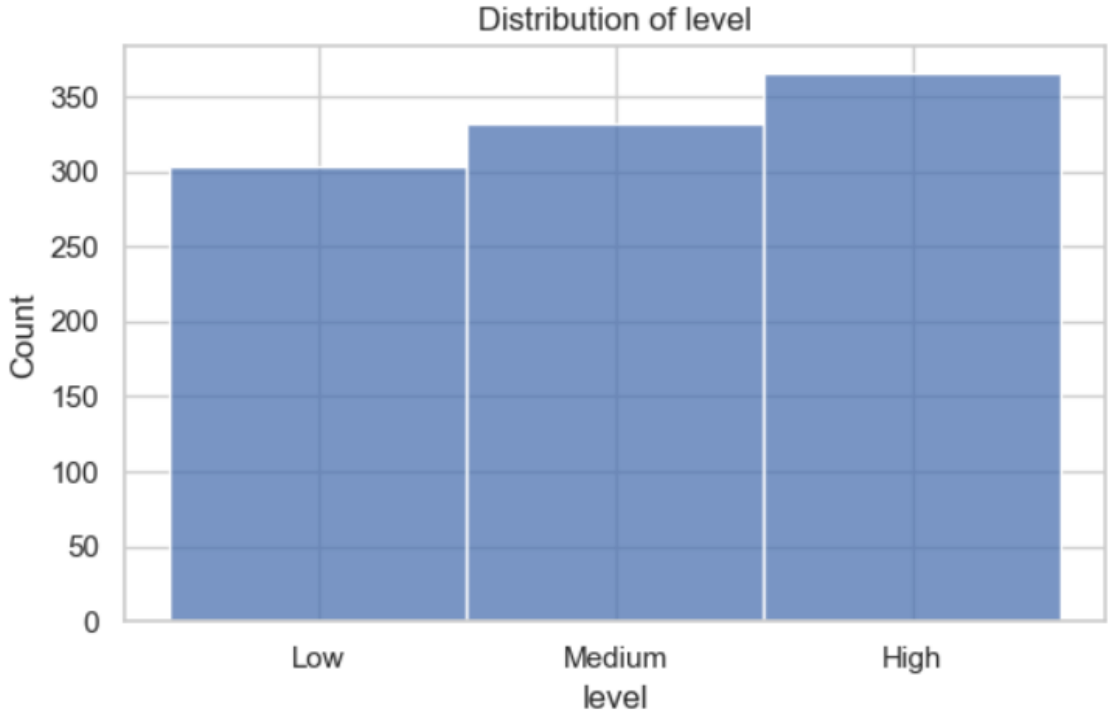

Distribution of swallowing_difficulty
Count
200
150
100
50
0
1
2
3
4
5
6
7
8
swallowing_difficulty


Distribution of clubbing_of_finger_nails
Count
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
clubbing_of_finger_nails

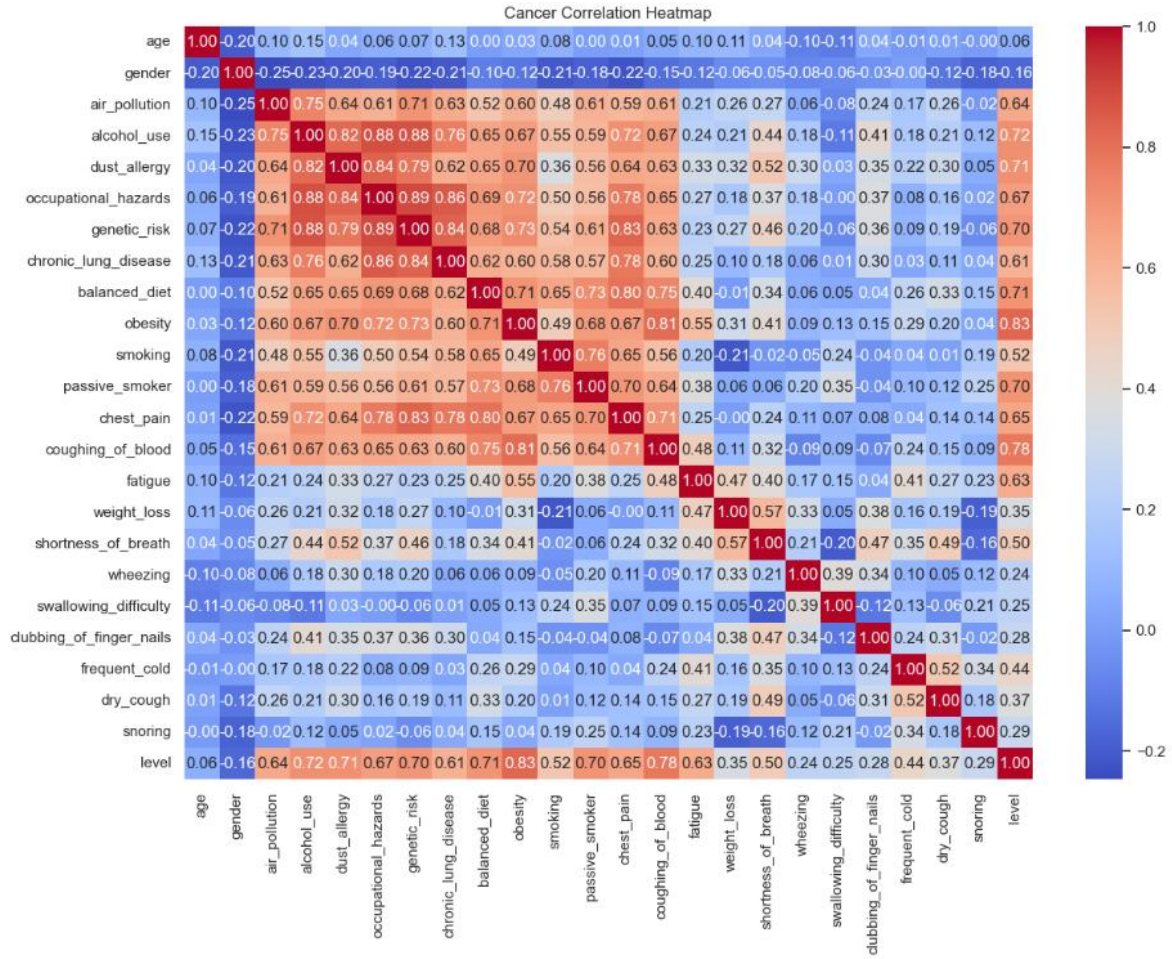

Distribution of frequent_cold
Count
200
150
100
50
0
1
2
3
4
5
6
7
frequent_cold


Distribution of dry_cough
Count
250
200
150
100
50
0
1
2
3
4
5
6
7
dry_cough


Distribution of snoring
Count
300
250
200
150
100
50
0
1
2
3
4
5
6
7
snoring


Distribution of level
Count
350
300
250
200
150
100
50
0
Low
Medium
High
level

The following Correlation heatmap was generated.



The following correlation coefficients were determined.

```
Correlation of features with target:

level                       1.000000
obesity                     0.827435
coughing_of_blood           0.782092
alcohol_use                 0.718710
dust_allergy                0.713839
balanced_diet               0.706273
passive_smoker              0.703594
genetic_risk                0.701303
occupational_hazards        0.673255
chest_pain                  0.645461
air_pollution               0.636038
fatigue                     0.625114
chronic_lung_disease        0.609971
smoking                     0.519530
shortness_of_breath         0.497024
frequent_cold               0.444017
dry_cough                   0.373968
weight_loss                 0.352738
snoring                     0.289366
clubbing_of_finger_nails    0.280063
swallowing_difficulty       0.249142
wheezing                    0.242794
age                         0.060048
gender                     -0.164985
Name: level, dtype: float64
```

The following Chi-scores were determined.

```
Chi2 Scores (Descending):
                     Feature  Chi2 Score
13         coughing_of_blood  658.892468
3                alcohol_use  620.887419
11             passive_smoker  609.497423
9                    obesity  578.975295
10                   smoking  517.156211
8              balanced_diet  478.645731
2              air_pollution  431.961151
12                chest_pain  399.411274
6               genetic_risk  394.776190
14                    fatigue  384.855493
5       occupational_hazards  330.540624
4               dust_allergy  328.886364
16       shortness_of_breath  275.210820
7       chronic_lung_disease  238.198406
19  clubbing_of_finger_nails  204.538310
15               weight_loss  173.532367
20             frequent_cold  170.516956
17                 wheezing  152.042346
21                dry_cough  118.636398
18     swallowing_difficulty   96.950451
22                   snoring   55.090080
0                        age   31.358688
1                     gender    2.021351
```

The following P-Values were determined.

```
P-values (Ascending):
                     Feature        p-value
13         coughing_of_blood  8.381438e-144
3                alcohol_use  1.499720e-135
11             passive_smoker  4.459809e-133
9                    obesity  1.892832e-126
10                   smoking  5.022912e-113
8              balanced_diet  1.157178e-104
2              air_pollution   1.587918e-94
12                chest_pain   1.857564e-87
6               genetic_risk   1.885557e-86
14                   fatigue   2.689624e-84
5       occupational_hazards   1.675003e-72
4               dust_allergy   3.830305e-72
16       shortness_of_breath   1.732725e-60
7       chronic_lung_disease   1.887441e-52
19  clubbing_of_finger_nails   3.846541e-45
15               weight_loss   2.079339e-38
20             frequent_cold   9.391038e-38
17                 wheezing   9.647704e-34
21                dry_cough   1.731544e-26
18     swallowing_difficulty   8.860884e-22
22                   snoring   1.089786e-12
0                        age   1.550770e-07
1                     gender   3.639730e-01
```

The following accuracy values were determined.

```
Accuracy (Random Forest): 1.0
Confusion Matrix RF:
[[61  0  0]
 [ 0 66  0]
 [ 0  0 73]]

Accuracy (XGBoost): 1.0
Confusion Matrix XGB:
[[61  0  0]
 [ 0 66  0]
 [ 0  0 73]]

RF Cross-Validation Accuracy: [1. 1. 1. 1. 1.]
XGB Cross-Validation Accuracy: [0.99 1.   1.   1.   1.  ]

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        61
           1       1.00      1.00      1.00        66
           2       1.00      1.00      1.00        73

    accuracy                           1.00       200
   macro avg       1.00      1.00      1.00       200
weighted avg       1.00      1.00      1.00       200

XGBoost Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        61
           1       1.00      1.00      1.00        66
           2       1.00      1.00      1.00        73

    accuracy                           1.00       200
   macro avg       1.00      1.00      1.00       200
weighted avg       1.00      1.00      1.00       200
```

Conclusion:

While the initial near 100% prediction accuracy was alarming, further investigation showed that this trend could also be a byproduct of the small population size, as there were only 1000 entries in the set.

Additionally, it can be determined that in this case, the specific strengths that apply to the random forest model make it better suited to this type of classification problem that XGBoost.

References

There are no sources in the current document.