Brice Derek Agnew (ST10072411)

IIE VARSITY COLLEGE  Cape Town

PDANA8411 POE PART 2

Contents

[Dataset and Models: 2](#_Toc199455206)

[Plan: 3](#_Toc199455207)

[Report: 5](#_Toc199455208)

[Conclusion: 23](#_Toc199455209)

[References 24](#_Toc199455210)

# 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](https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link)] 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.

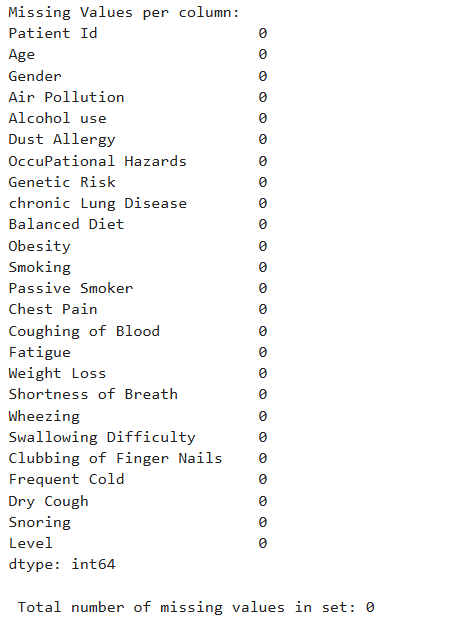
* Exploratory Data Analysis: Key statistics and distributions will be identified and presented using visual representations.

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.

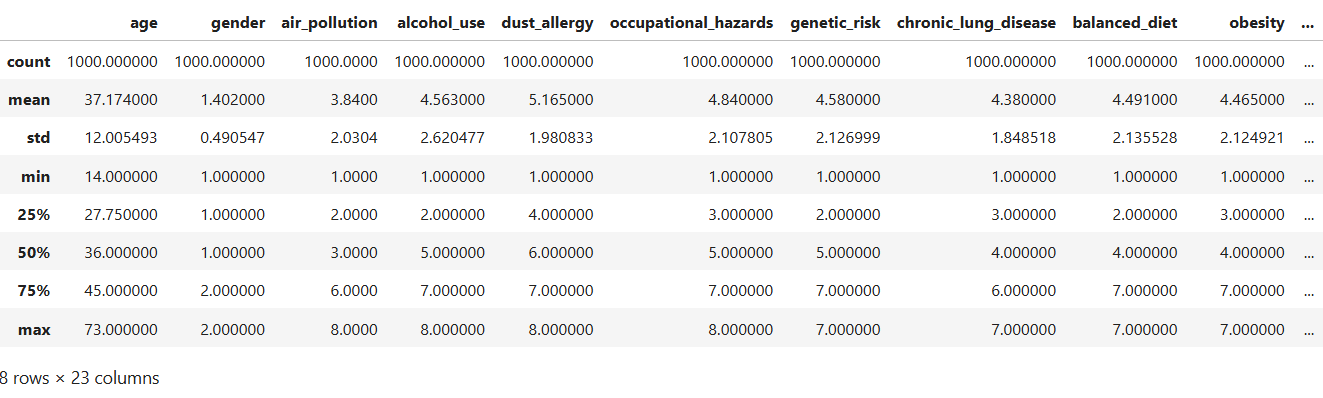


It was found that there were no duplicate columns or rows, and no further action was needed.

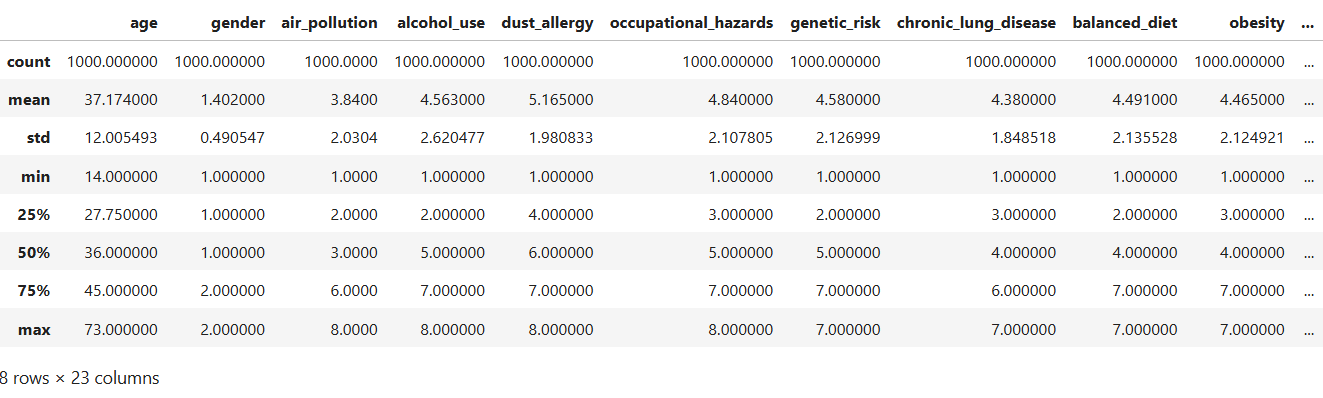
A white background with black text

AI-generated content may be incorrect.

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



Key statistics for each feature were identified.



The data visualizations for each of the features are as follows:  
  
A graph of age distribution

AI-generated content may be incorrect.

A graph with a bar

AI-generated content may be incorrect.

A graph of air pollution

AI-generated content may be incorrect.

A bar graph with blue bars

AI-generated content may be incorrect.

A graph with blue bars

AI-generated content may be incorrect.

A graph of hazard

AI-generated content may be incorrect.

A graph with blue bars

AI-generated content may be incorrect.

A graph of a number of patients

AI-generated content may be incorrect.

A graph of a weight loss

AI-generated content may be incorrect. A graph with blue bars

AI-generated content may be incorrect. A graph of smoking distribution

AI-generated content may be incorrect. A graph of a distribution of smoker

AI-generated content may be incorrect. A graph of a pain

AI-generated content may be incorrect. A graph of a disease

AI-generated content may be incorrect. A graph of fatigue

AI-generated content may be incorrect. A graph of weight loss

AI-generated content may be incorrect. A graph with blue bars

AI-generated content may be incorrect.

A graph of a distribution of wheezing

AI-generated content may be incorrect. A graph with blue bars

AI-generated content may be incorrect. A graph of a bar

AI-generated content may be incorrect. A graph of a number of blue bars

AI-generated content may be incorrect. A graph of a distribution of dry cough

AI-generated content may be incorrect. A graph of a distribution of snoring

AI-generated content may be incorrect. A diagram of a distribution of level

AI-generated content may be incorrect.

The following Correlation heatmap was generated.

A screenshot of a chart

AI-generated content may be incorrect.

The following correlation coefficients were determined.  
A screenshot of a computer

AI-generated content may be incorrect.

The following Chi-scores were determined.

A screenshot of a computer

AI-generated content may be incorrect.

The following P-Values were determined.

A screenshot of a computer

AI-generated content may be incorrect.

The following accuracy values were determined.

A screenshot of a computer

AI-generated content may be incorrect.

# 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.**