1st January 2024

Kieran Harley Student Number: S2214057

“I declare that all work submitted for this coursework is the work of Kieran Harley alone unless stated otherwise.”

Machine Learning & Data Analytics

Coursework Glasgow Caledonian University

Table of Contents

[Table of Figures 1](#_Toc155039187)

[Introduction 2](#_Toc155039188)

[Problem Definition 2](#_Toc155039189)

[Dataset 2](#_Toc155039190)

[Construction of Machine Learning Model 3](#_Toc155039191)

[Support Vector Machine (SVM) Learning Model 3](#_Toc155039192)

[Data Exploration 3](#_Toc155039193)

[Data Pre-Processing 5](#_Toc155039194)

[Training the Support Vector Machine 6](#_Toc155039195)

[Testing 7](#_Toc155039196)

[Test Results 7](#_Toc155039197)

[Discussion 8](#_Toc155039198)

[Significance of Data Pre-Processing 8](#_Toc155039199)

[Effectiveness of Model Tuning 8](#_Toc155039200)

[Performance of Support Vector Machine on Dataset 8](#_Toc155039201)

[Reference List 9](#_Toc155039202)

# Table of Figures

[Figure 1: Dataset Column Headings 2](#_Toc155039174)

[Figure 2: Python Imports 3](#_Toc155039175)

[Figure 3: Python Data Frame Head 3](#_Toc155039176)

[Figure 4: Data Frame Types 4](#_Toc155039177)

[Figure 5: Data Frame Count Missing Values 4](#_Toc155039178)

[Figure 6: Splitting Data into Target and Input Variables 5](#_Toc155039179)

[Figure 7: One Hot Encoding 5](#_Toc155039180)

[Figure 8: Split Data into Training and Validation 6](#_Toc155039181)

[Figure 9: Accuracy of Validation Model 6](#_Toc155039182)

[Figure 10: Accuracy of Test Model 7](#_Toc155039183)

[Figure 11: Test Result 7](#_Toc155039184)

# Introduction

This report details how to a csv file can be interpreted with modern machine learning methods, to provide insights into pattern recognition and data analysis.

## Problem Definition

To develop a predictive model capable of assessing a dataset pertaining to a study on lung cancer. The model will interpret relationships between attributes taken from a dataset focussing on common symptoms, with an aim of predicting the likelihood of a test case developing lung cancer. By constructing a robust machine learning model, the goal is to aid early detection of lung cancer and aid the medical industry by giving another tool to healthcare professionals, working in diagnostics.

## Dataset

The csv file selected for this investigation was populated with information from a Survey on Lung Cancer, and the dataset itself was found on the Kaggle website. This dataset was chosen as it contained sixteen common attributes, over three-hundred-and-nine instances, making this a good sample size to investigate. Another thing that contributed to choosing this dataset was that there were no missing values, making it a very healthy sample to explore.

The attributes within this dataset pertain to a survey questionnaire. The attributes are listed as: gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath difficulties swallowing, chest pain and finally – lung cancer.

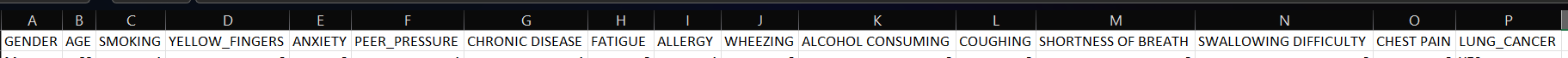


Figure 1: Dataset Column Headings

These attributes are an aggregation of elements that are commonly associated with lung cancer. With this information, a supervised learning model can be used to assess the commonality and then produce an accurate assessment of the correlation between the likelihood of lung cancer, against these attributes.

To make sure this information is effectively interpreted, choosing the correct method of machine learning is paramount. In this case a supervised learning model is the most appropriate as this dataset is labelled and the aim is to give a prediction at the end. Furthermore, the right kind of supervised learning model is also important to ensure the best results. As there is a healthy amount of data and the outcome is to find a classification between multiple attributes, the support vector machine (SVM) was chosen.

# Construction of Machine Learning Model

## Support Vector Machine (SVM) Learning Model

The SVM machine learning model is a powerful tool for classification and regression analysis. Widely used for effectiveness in handling both linear and non-linear relationships in data modelling, the SVM will find an optimal hyperplane, separating different classes.

As SVMs are inherently binary classifiers, this method of supervised learning algorithm is well suited to the data from the lung cancer survey, as this can classify two responses into two categories. One example of this is the result of the patient ultimately having lung cancer or not.

## Data Exploration

Prior to pre-processing, it is always a good idea to do preliminary data exploration. This is because, by exploring the data the coder will have a better understanding of which kind of data types are available in the dataset. Furthermore, this can also highlight any potential missing values.

Making use of the “Pandas” library within Python, this csv file can be prepared for exploration. Loading the data in this method allows developers to load the file into the integrated development environment by assigning it to a variable. In this case – ‘df’ (data frame).

A screen shot of a computer program

Description automatically generated

Figure 2: Python Imports

Discovering the shape of the model, allows for greater overview of the dimensions within. This is called upon by using the dot operand and calling the ‘shape’ (shown below). This returns the number of rows and columns within the dataset. Another method is to show a few rows of the data frame. Again, this uses the dot operand and ‘head’ command. The head command is good for inspecting the content of the data prior to manipulation.

A screenshot of a computer

Description automatically generated

Figure 3: Python Data Frame Head

Another method of data exploration in this file is the use of the ‘info’ command. This gives overview of the datatypes within the file. This information is useful when it comes to debugging.

A screenshot of a computer

Description automatically generated

Figure 4: Data Frame Types

The last method of data exploration used for this algorithm is to check the number of null cells within the file. Null cells can represent false values, thus placing importance on discovery prior to data preprocessing to ensure the most accurate results.

A screenshot of a computer

Description automatically generated

Figure 5: Data Frame Count Missing Values

## Data Pre-Processing

Ensuring that the raw data is correctly prepared is pivotal in the process of data analytics. This process is called data pre-processing, and it is responsible for making the raw data into a more readable format. In this SVM model, this is implemented by separating the target variable from the rest of the input variables. This is done by ‘dropping’ the input variable to ensure the data within this column is not considered at runtime and ensuring the model generalises well to new unseen data.

A screenshot of a computer program

Description automatically generated

Figure 6: Splitting Data into Target and Input Variables

Thanks to prior data exploration, it was clear that there is an object type within the datafile which could break the SVM. Further research was conducted to find a way to include the column of ‘GENDER’, as this could be used in later examples to further drill down on differences between men and women’s likelihood of developing this disease, with the same input variables.

After research had concluded, it was clear that the ‘one hot’ method was key for including this data, as this encodes the M/F (male / female) from the dataset into a binary field, meaning that the value is now represented by 1’s and 0’s which allows it to be included within the training set.

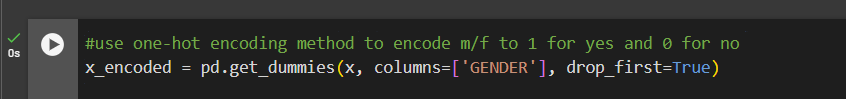


Figure 7: One Hot Encoding

The next step in data pre-processing is to split the data set into training data into training and validation.

A screenshot of a computer program

Description automatically generated

Figure 8: Split Data into Training and Validation

The training data exists for the model to be able to make future predictions against the dataset. The validation set exists for the model to be able to assess how well the trained model can generalise new data which aids to prevent overfitting. This division of data is fundamental in creating a robust and effective learning algorithm.

## Training the Support Vector Machine

How well an SVM can perform as a data classifier and conduct pattern recognition is down to well the algorithm has been trained. Ensuring that an SVM is trained properly is incredibly important as an improperly trained SVM can lead to skewed results, obfuscating the end figure.

This classifier was trained by making a prediction against the x & y training variables to show a working metric through validation. The model then shows an accuracy score to aid reliability in its predictions. The model wants to be as close to ‘1’ as possible as this would represent 100% perfect score prediction.

A screen shot of a computer program

Description automatically generated

Figure 9: Accuracy of Validation Model

The printed statement indicates an exceptionally high percentage Validating the result gives greater overview of how the model will generalise new information.

# Testing

After validation it is equally important to check the score of the test data, as this can highlight whether the learning model has issues such as overfitting or underfitting when analysing the dataset.

The following code was added to the algorithm to check the accuracy of the machine learning model against the test dataset.

A screenshot of a computer program

Description automatically generated

Figure 10: Accuracy of Test Model

By using clf1, this code can predict the target values of ‘y\_test\_pred’ against the input of ‘x\_test’, before then giving an accuracy score of ‘y\_test’ and ‘y\_test\_pred’ using the metrics function from the scikit-learn Python library.

## Test Results

From this code, the output against the dataset when testing the machine learning model is 87.23%.



Figure 11: Test Result

As this score is high like the validation score, the chances of any overfitting or underfitting issues are minimal and the SVM can be considered a success.

In conclusion, it can be assumed that against the previously listed symptoms of lung cancer, this SVM machine learning model can predict the presence or absence of lung cancer with a high accuracy percentage and perform well against new unseen data.

# Discussion

## Significance of Data Pre-Processing

Data pre-processing plays a huge role in the performance of an SVM. Preprocessing ensures that the input data is appropriately formatted and standardised which enhances the SVMs ability to recognise patterns. Some vital steps to pre-processing data include, handling missing values, scaling features, and encoding variables (Kang and Tian, 2018). Proper pre-processing aids SVMs in working faster during training, improving efficiency, and mitigating any issues caused by feature scaling. Furthermore, this process helps to identify relevant patterns which help to optimise the model’s accuracy and making the model better at generalising new data.

Overall, meticulous data pre-processing positively impacts the result of the SVM by refining the input data for more effective machine learning.

## Effectiveness of Model Tuning

The efficacy of a SVM as a data classifier and for a pattern recognition relies heavily on how it is trained and tuned. Ensuring proper training is paramount as an inadequately trained SVM can yield biased results and distorting the outcome. This is why optimising features such as hyperparameters is very important to ensuring a solid machine learning algorithm (Pannakkong et al., 2022).

The SVM model highlighted in this document underwent training by predicting against the x and y training variables, establishing a working metric which was validated through model accuracy scores. The model accuracy score is a good way of displaying how the model performed and gives greater overview of reliability.

## Performance of Support Vector Machine on Dataset

As shown in figures 9, 10 and 11, this machine learning algorithm has a high rate of success in classifying the data within the dataset. One of the reasons that the classifier can perform so efficiently is because the most optimal algorithm was selected for this type of data. As the source data is high dimensional, relative to the type of dataset, a support vector machine was a clear choice for data analysis.

Because of this, the algorithm can perform well at the task of rationalising connections between the common symptoms of lung cancer and the presence of the disease.

# Reference List

Kang, M. and Tian, J. (2018). Machine Learning: Data Pre-processing. *Prognostics and Health Management of Electronics*, pp.111–130. doi:https://doi.org/10.1002/9781119515326.ch5.

Pannakkong, W., Thiwa-Anont, K., Singthong, K., Parthanadee, P. and Buddhakulsomsiri, J. (2022). Hyperparameter Tuning of Machine Learning Algorithms Using Response Surface Methodology: A Case Study of ANN, SVM, and DBN. *Mathematical Problems in Engineering*, 2022, pp.1–17. doi:https://doi.org/10.1155/2022/8513719.