# CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

## 3-A – Methodology

Purpose of the project is to create a machine learning program for finding the factors of lung cancer and predicting how it affects to patients, and finally detecting the level of lung cancer of the given patient. Data is pre processed in order to apply certain operations on it. We started by analysing which of the conditions given in the dataset affects the cancer level. After we determined those conditions, we predicted the level of cancer for given patient through the machine learning model we created.

## 3-B – Conjectures and Exploratory Questions

Cancer is a complex group of diseases with many possible causes. Signs and symptoms caused by cancer will vary depending on what part of the body is affected. In our case we will examine lung cancer. General symptoms and causes for lung cancer is smoking, fatigue, persistent cough and trouble breathing. So our guess is that we will see increase in cancer level with patients who have this symptoms or habits.

## 3-C. Research Procedures

### a) Epistemology

To create a successful machine learning program, we started by analysing the data by hand and examined it's structure. After we learned what data provides us and possible theories we discover from this data, we started to write our computer program. We used classification and clustering algorithms to select important features and display relevant graphs to create a solid outcome. Then we tested our programs prediction accuracy by comparing it to actual results.

### b) Theoretical Perspective

Lung cancer is dangerous and complicated. Our project could potentially allow us to have a clearer and predictable view of this disease. With enough data, we can predict what kind of lifestyle and symptoms have effect on lung cancer and how much they affect, with high precision. This will help patients to detect or prevent their disease at an early stage, which is very crucial for cancer. Also, because it is a machine learning program, eventually there could be a service for people to use this program on their own to see how much risk they are in without going to the doctor. We know that machine learning programs always keep improving as they are used on larger scales, so these kinds of improvements are not impossible and could help many people.

### c) Methodology

Our method is to analyse the data with a machine learning program specifically designed for this problem. We selected to use machine learning because while humans or normal programs get slower and less accurate as data grows, machine learning algorithms only get better and more accurate. Our data contains lost of columns and rows, so choosing machine learning to make predictions and completing our task was the best choice.

### d) Used Methods

Our program is written in the Python programming language. For machine learning algorithms we used libraries such as Sklearn, Keras, Tensorflow. To operate on data we used NumPy, Matplotlib and Pandas. We imported the data using OS library by importing it as .xlsx file and converting it to .csv. We read the data by using pandas and used matplotlib to create figures. We used SelectKBest algorithm for feature

selecting(finding out features that have the most effect) which is the foundation of our program. We used sklearn library to train and test our data. For modelling we used the support vector machines classifiers (SVC).

**e) Coding system**

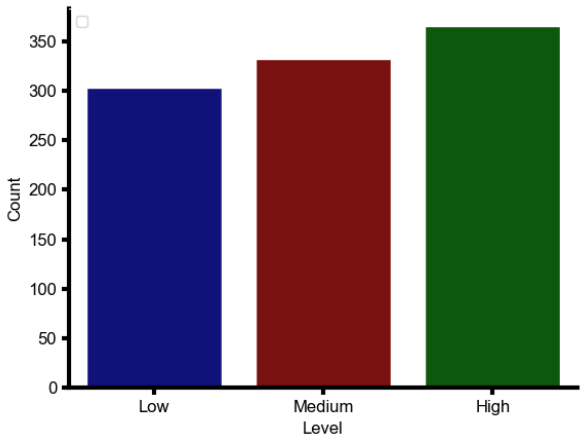
Our program starts by loading the data, then we check if the data has any problems or unbalanced. We first checked if gender has any correlation between cancer level. Then we further break it down by also analysing alcohol use and fatigue effects by gender. Then we analysed if there are any correlation of age. After that we used SelectKBest Algorithm to find out most important features. We created a scoring system (from 0 to 1200 according to their importance) for all of the features and selected the ones that scored higher than 200. Than we split the data and trained a scaler model to apply to a data set. ¼ of the data is used for testing while ¾ of it used for training. For modelling we used support vector machines classifiers (SVC). After that we checked our programs accuracy by comparing it's Grid Search and Random Search algorithm results to actual results.

### 3-D – Human Participants and Ethics Precautions

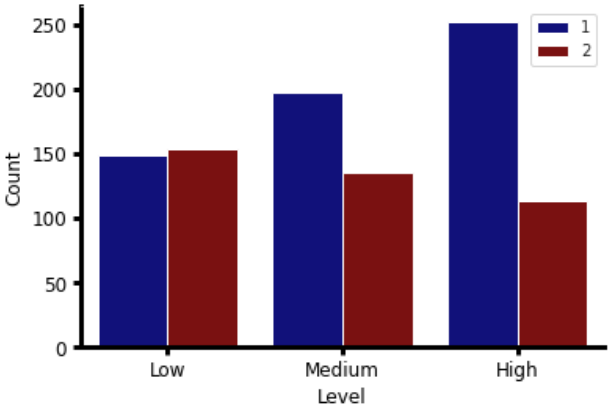
The data we are using does not contain and personal information about the human participants. Only ethical problem could be that these informations could be taken without their permission which could be avoided by getting information from trusted resources when building the final program. Because the data we used doesn't contain any personal information or any information that could harm any person or cause privacy issues, our only concern is the will of the people who give their information.

## CHAPTER 4: RESULTS AND FINDINGS

After implementing the methods explained in the previous chapter, we did get some interesting findings. We observed the effect of gender, we created a score system to see the importance of all symptoms and lifestyle. Our results of prediction was %100 accurate. Reason for %100 accuracy is we trained and tested our program on the same dataset which results in very high accuracies in consistent dataset. Because of this data is very consistent our accuracy was full on point. Still, this doesn't mean that our programs accuracy will be significantly low after we test it with different combinations of different datasets. One of the important things we discovered is that graphs could be deceiving if only shown without explanation or a perspective from the bigger picture. Detailed explanation is given below the graphs below.

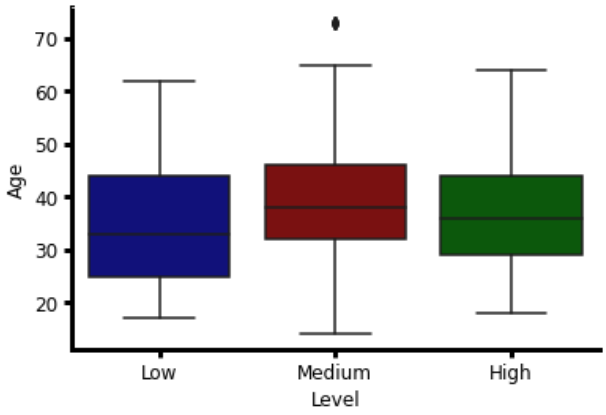


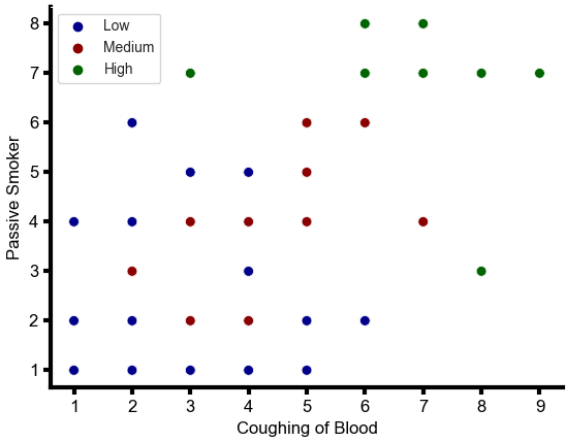

First graph is not relevant to the exact results, but very important for our programs accuracy. This graph compares the balance of 3 level of cancers in the data. As we can see data is not unbalanced and could be used to construct machine learning programs. If data was not balanced, results could not be trusted to make accurate predictions.



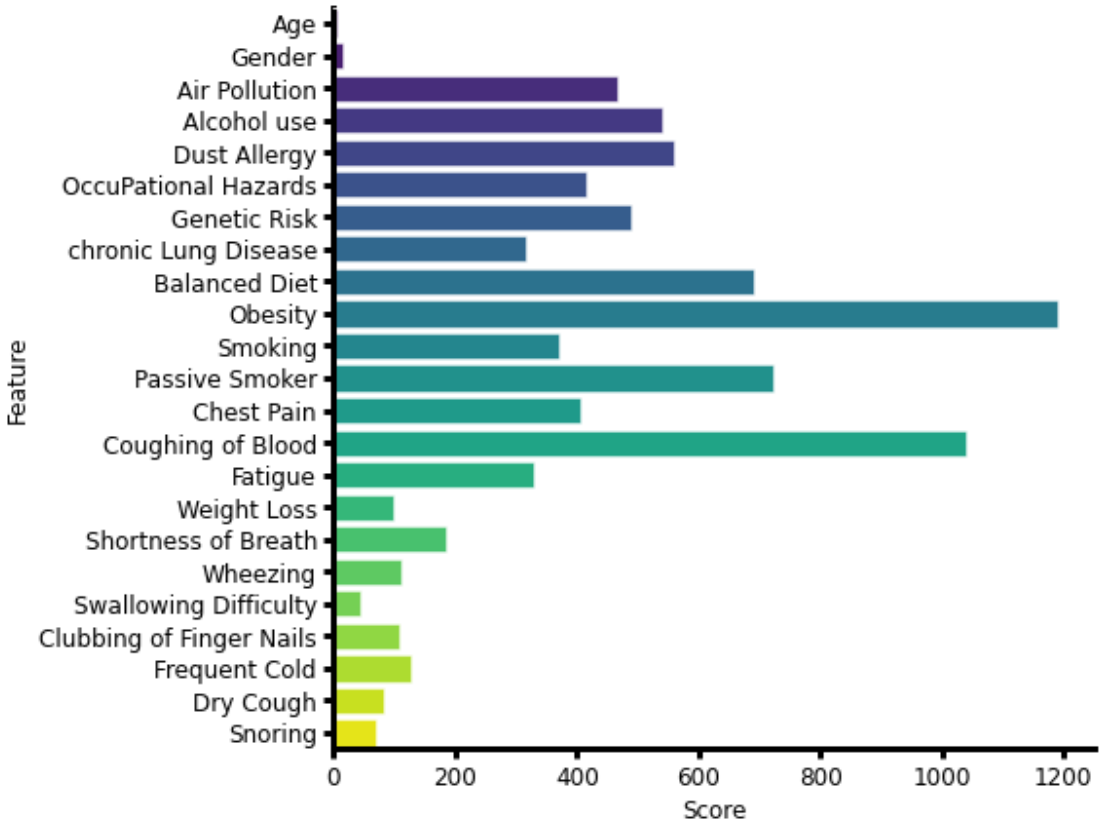

Graph on the left shows the relation between three cancer levels (low, medium, high) and genders(blue: male, red: female). As we can see from this graph, while first two phases of the cancer is almost equally distributed between males and females, there is a huge gap between those two genders in the third level of cancer. We can see that gap is starting to increase after medium level and reaches it's maximum at high level

Graph on the left show the relation between age of patients and their cancer level. As we can see, there is no visible affect of age in the level of cancer.

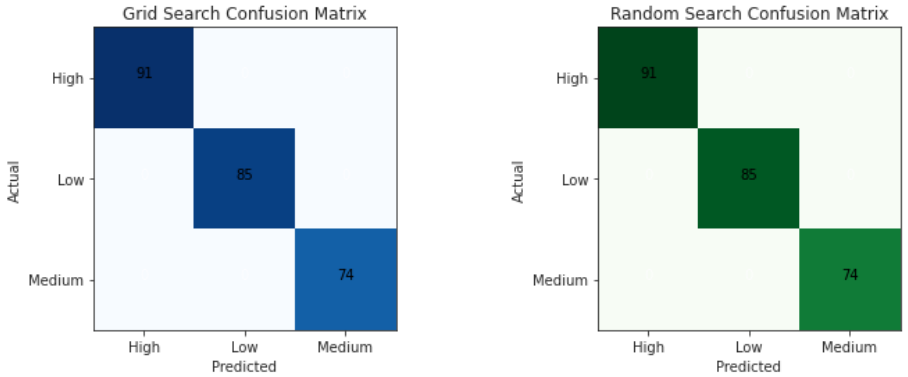



Graph on the left shows one symptom(Coughing of Blood) and one lifestyle feature(Passive Smoking) and their effect on cancer level. As numbers from 1-9 increases, severity of that feature increases. We can clearly see some clusters forming especially in severe cases. Comparison and correlation of two or more features can be graphed like this. We choosed 2 of the most effective features for example but every aspect can be compared and examined.




Graph above is the visualization of our feature selection algorithms results. As we can see, the graph contains all of the features. Score in the x axis shows the importance of that feature. This graph is a little bit tricky because it may deceive people who look at it for the first time. In the graph we see that obesity is playing a huge part in the level of the cancer. We were surprised from this result and after doing some research, we saw that (according to: Obesity and incidence of lung cancer: A meta-analysis from IJC) overweight and obesity are protective factors against lung cancer, especially in current and former smokers. So we can't just simply say that obesity plays a huge factor because it doesn't. Our data is big, but still not big enough for some things to occur out of luck. Other than this all of the other factors are in place and we can use them to construct our algorithm.




Our final graph is showing accuracy of our program by using SVC Grid and SVC Random. At y axis there are the actual values and at x axis there are the predicted values for three levels of cancer. As we can see, our programs accuracy for this dataset is %100 percent in both types of SVC's.