Applying Gradient Boosted Decision Tree Algorithm for Predicting Interstitial Lung Cancer

Authors: Mohammad Nafees Bin Zaman, Ahmed Amer Shafie Abdelrahman Aly, MHD Khaled Maen, Mahamat Nour Ali Mai, Ilham Fadhillah Amka

Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia.

zamannafees@gmail.com, ahmed3aamer96@gmail.com, mk.maen93@gmail.com, mahamatnouralimai@gmail.com, ilhamfadhillahamka1@gmail.com

Author Note

This report holds all the information about the group project of Machine Learning

Matric no: 1616357, 1432285,1523591,1510455,1434071

Group Number: 08

Date: 12/12/2018

Submitted to: Dr. Amelia

Table of Content

Abstract: According to the World Health Organization (WHO), 7.6 million deaths globally each year are caused by cancer; cancer represents 13% of all global deaths. Among all types of cancer, lung cancer is by far the number one cancer killer. Early detection of such a disease may save thousands of peoples live as well as their wealth. This paper illustrates our work of detecting the possibilities of having Lung Cancer, using a Machine Learning Algorithm. We use Gradient Boosted Decision Tree since it's an interpretable and optimized algorithm which gives us a good accuracy for prediction.

Keywords**:** Lung Cancer, Gradient Boosted Decision Tree, Prediction etc.

Introduction: In this modern era, we all are dependent on machines. Almost every single act in our everyday life is operated by different machines. As a result, it is impossible to think a single day without even having a simple smart phone, which explains the current scenario of this technologically advanced world. However, one may ask how these devices have so much influence on our daily life activities. To answer this question, we simply can refer to machine learning. Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions. The extraordinary success of machine learning has made it the default method of choice for AI researchers and experts. Indeed, machine learning is now so popular that it has effectively become synonymous with artificial intelligence itself. As a result, it's not possible to tease out the implications of AI without understanding how machine learning work as well as how it doesn't. That is why it has so much effect on our daily life. As machine learning has proven its importance and necessity, machine learning in medical science is no other exception. Our project is harmonizing medical science with computer science as we are predicting lung cancer through machine learning algorithm which is Gradient Boosted Decision Tree.

Applying Gradient Boosted Decision Tree Algorithm for Predicting

Interstitial Lung Cancer

Objectives

- Collect information and dataset for interstitial lung cancer with relevant label.
- Analyze and visualize dataset to predict the chance of lung cancer correctly.
- Develop a machine learning model using Gradient Boosted Decision Tree and implement in real life scenarios.
- Conclude the result based on the outcome of our constructed algorithm.
- Measure how much smoking affect lung cancer.

Expected Outcomes

Successfully build a machine learning model using Decision Tree which can predict the chances of having Lung Cancer whether it's low, medium, or high.

Literature Review

Related Work

It is a great gesture of the scientists that they are trying to harmonize lot of areas to make things easier for human being. Medical science and Computer Science is a great example of it. Our research work is also inspired by these previous works which is done by so many researchers. Mainly in this case, predicting lung cancer most of the researchers chose bagging and boosting algorithm [3] whereas a few numbers of researchers took decision tree [5]. So, we choose to work with the decision tree and to improve it for further purposes. All the related works are cited in the references section.

Lung Cancer

Cancer, also called malignancy, is abnormal growth of cells. There are more than 100 types of cancer, including skin cancer, breast cancer, lung cancer, lymphoma etc. Symptoms of these cancers vary depending on the type. Several types of cancer treatment are available including chemotherapy, radiation, and/or surgery but none of them can assured fully to cure this disease. Lungs are two spongy organs in chest that take in oxygen when human inhales and release carbon dioxide when exhales. One of the most dangerous type of cancer is lung cancer that attacks the lungs. It is the leading cause of cancer deaths in Malaysia, among both men and women. Lung cancer claims more lives each year than do colon, prostate, ovarian and breast cancers combined.

Signs and Symptoms

Lung cancer usually doesn't show signs and symptoms in its earlier stages. It typically occurs only when the disease is grown up. Signs and symptoms of lung cancer may include:

- A new cough doesn't go away.
- Coughing up blood, even small amount.

- Shortness of breath.
- Chest pain.
- Hoarseness.
- Losing weight without trying.
- Bone pain.
- Headache.

Risk factors

A number of factors may increase the risk of lung cancer. Some risk factors can be controlled, for instance, by quitting smoking. And other factors can't be controlled, such as your family history.

Risk factors for lung cancer include:

- **Smoking:** Your risk of lung cancer increases with the number of cigarettes you smoke each day and the number of years you have smoked. Quitting at any age can significantly lower your risk of developing lung cancer.
- **Exposure to secondhand smoke:** Even if you don't smoke, your risk of lung cancer increases if you're exposed to secondhand smoke.
- **Exposure to radon gas:** Radon is produced by the natural breakdown of uranium in soil, rock and water that eventually becomes part of the air you breathe. Unsafe levels of radon can accumulate in any building, including homes.
- **Exposure to asbestos and other carcinogens:** Workplace exposure to asbestos and other substances known to cause cancer such as arsenic, chromium and nickel also can increase your risk of developing lung cancer, especially if you're a smoker.
- **Family history of lung cancer:** People with a parent, sibling or child with lung cancer have an increased risk of the disease.
-

Prevention

There's no sure way to prevent lung cancer, but you can reduce your risk if you:

- **Don't smoke.** If you've never smoked, don't start. Talk to your children about not smoking so that they can understand how to avoid this major risk factor for lung cancer. Begin conversations about the dangers of smoking with your children early so that they know how to react to peer pressure.
- **Stop smoking.** Stop smoking now. Quitting reduces your risk of lung cancer, even if you've smoked for years. Talk to your doctor about strategies and stop-smoking aids that can help you quit. Options include nicotine replacement products, medications and support groups.
- **Avoid secondhand smoke.** If you live or work with a smoker, urge him or her to quit. At the very least, ask him or her to smoke outside. Avoid areas where people smoke, such as bars and restaurants, and seek out smoke-free options.

- **Test your home for radon.** Have the radon levels in your home checked, especially if you live in an area where radon is known to be a problem. High radon levels can be remedied to make your home safer. For information on radon testing, contact your local department of public health or a local chapter of the American Lung Association.
- **Avoid carcinogens at work.** Take precautions to protect yourself from exposure to toxic chemicals at work. Follow your employer's precautions. For instance, if you're given a face mask for protection, always wear it. Ask your doctor what more you can do to protect yourself at work. Your risk of lung damage from workplace carcinogens increases if you smoke.
- **Eat a diet full of fruits and vegetables.** Choose a healthy diet with a variety of fruits and vegetables. Food sources of vitamins and nutrients are best. Avoid taking large doses of vitamins in pill form, as they may be harmful. For instance, researchers hoping to reduce the risk of lung cancer in heavy smokers gave them beta carotene supplements. Results showed the supplements actually increased the risk of cancer in smokers.
- **Exercise most days of the week.** If you don't exercise regularly, start out slowly. Try to exercise most days of the week.

Lung cancer in Malaysia

According to the 2014 World Health Organization report, lung cancer accounted for 19.1 deaths per 100,000 population in Malaysia or 4,088 deaths per year (3.22% of all deaths), the second most common cause of death due to cancer in the country after breast cancer, and the eighth most common cause of death from all causes. In 2014, cancer of the trachea, bronchus and lung accounted for 24.6% of all cancer mortality in males in the country, the most common cancer death, while in females, it accounted for 13% of all cancer deaths, the second most common cancer death after breast cancer. 4,403 lung cancers were diagnosed in the country in 2014, 3,240 in males (the most common cancer diagnosed), and 1,163 in females (the fourth most common cancer diagnosed). Information on the epidemiology of lung cancer was also obtained from the National Cancer Registry (NCR). From its last published report in 2007, lung cancer was the third most common cancer in

the country, the second most common cancer in males and the 4th most common in females. The mean age at which lung cancer is diagnosed in Malaysia is about 60 years with a peak age of diagnosis in the 7th decade. The incidence of diagnosed lung cancer in Malaysian patients aged less than 40 years is relatively low at approximately 6.2%. [1]

Machine Learning Model

In our project we apply Decision Tree algorithm to achieve our goal from the selected dataset. The Decision Tree algorithm builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf node. The core algorithm for building decision trees called *ID3* by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses Entropy and Information Gain to construct a decision tree.

- Calculate the entropy using frequency tables.
- Information Gain: The information gain is based on the decrease in entropy after a dataset is split on an attribute.
- Calculate entropy of the target.
- The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain or decrease in entropy.
- Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.
- A branch with entropy of 0 is a leaf node.
- A branch with entropy more than 0 needs further splitting.
- The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

Experimental Setup

We follow the data science process to analyze the problem and to complete our project successfully [Figure 1]

Data Collection

The dataset is about lung cancer patients in US. it consists of 1000 observations to people who have different lung diseases. There are total of 25 columns (attributes); patient id, 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 and level. The level attribute is the label or dependable variable of the dataset. It has three values low, medium and high and it is labelled based on the features mentioned above. So, the scenario of the dataset is a supervised classification problem.

Dataset:

https://drive.google.com/drive/folders/1lKdvbjxle_9jU5tNZV9kjDRvyyitGuit?usp=sharing

Data Processing

We started to explore further the dataset with a goal of finding any issue and then apply data engineering method to prepare the dataset for our model. By running basic R functions such as “str()”, “ summary()”, and “head()” we found that our dataset is almost clean since there are no outliers nor missing values. Furthermore, there are no any duplicated rows Observations are distributed in a balanced way according to the three labels which ensure our model can learn all classes. The only issue we found is related to metadata; most of variables are considered integer however they are not continuous values. Considering

decision trees can handle categorical data easier than numeric data we converted all of those variables to factors except the "age" variable which should be integer as it is.

```
str(cancer_data)
head(cancer_data,1)
summary(cancer_data)
variables <- 3:24
for(i in variables){
 cancer_data[,i] <- factor(cancer_data[,i])
}
```

Feature Importance

We check feature importance in order to try avoiding some noisy features. We used "filterVarImp" function from "caret" package which compare each variable with the label and measure their influencing degree. We found that Features can be divided into two parts; Syndromes and Causes. The most important cause of lung cancer depending on our data is alcohol using. Other causes heavily affected includes; passive smoking, obesity, air pollution, and genetic reasons. Age, and Gender doesn't influence results that much which may be removed from model training later. The most significant Syndromes to include coughing blood, fatigue, and chest pain. [Figure 4]

To be more precise on data processing and cleaning we used azure machine learning studio as well. [Figure 2]

Exploratory Data Analysis (EDA)

After cleaning the data, we plot graphs such as histograms and scatter diagrams as well as a statistical summary to help understand more about the correlations and relationships between the variables. Furthermore, this helps to identify any outliers that were missed during data cleaning. If anomalies are identified, data is further cleaned, and EDA is reconstructed again. [Figure 3a, Figure 3b]

Machine Learning Algorithm

At this part we do the main business where we use Gradient Boosted Decision Tree algorithm to predict Lung Cancer. Our dataset consists of 25 variables where 24 of them are independent variable and the rest one is the only dependable variable by which we or our model have to predict either a patient has a possibility of having cancer or not. We train our model by feeding data from the dataset and get an accuracy of 91.4 % of our model. It means our model can predict possibilities more than 91 percent accurately.

Data Visualization

Data visualization is one of the effective ways to interact with the data. Results are interpreted and visualized to report our findings, which is used to communicate with real life problems, in our case which is Lung Cancer. Data visualization can easily be interpreted and understood by everyone. So, visualization of data from our research and trained model could also help the doctors to give proper guidance to the patients. Furthermore, patients could consult with the doctor also by seeing his/her chance of having lung cancer. [Figure 5]

Data Product

Data product is a commercial based model that can be used in general and commercially. In our case, as it is as experimental process, we do not prefer our model as a data product. We recommend trying other models that are available.

Feedback Loop

New hospitals will eventually get involved with this model once we publish it as a data product, thus generating more data that can be fed into the model to further refine it and help to make it more accurate.

Calculation:

Gradient Boosted Decision Tree is an optimized algorithm which uses multiple decision trees. A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous).

Entropy

ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one. To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

a) Entropy using the frequency table of one attribute:

Entropy (Set) =

$$\sum_{i=1}^{C} -p_i \log_2 p_i$$

b) Entropy using the frequency table of two attributes:

E (T, X) =

$$\sum_{C \in x} P(c) E(c)$$

Information Gain

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain.

Step 1:

Calculate entropy of the target.

Step 2:

The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain or decrease in entropy.

Step 3:

Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

Step 4a:

A branch with entropy of 0 is a leaf node.

Step 4b:

A branch with entropy more than 0 needs further splitting.

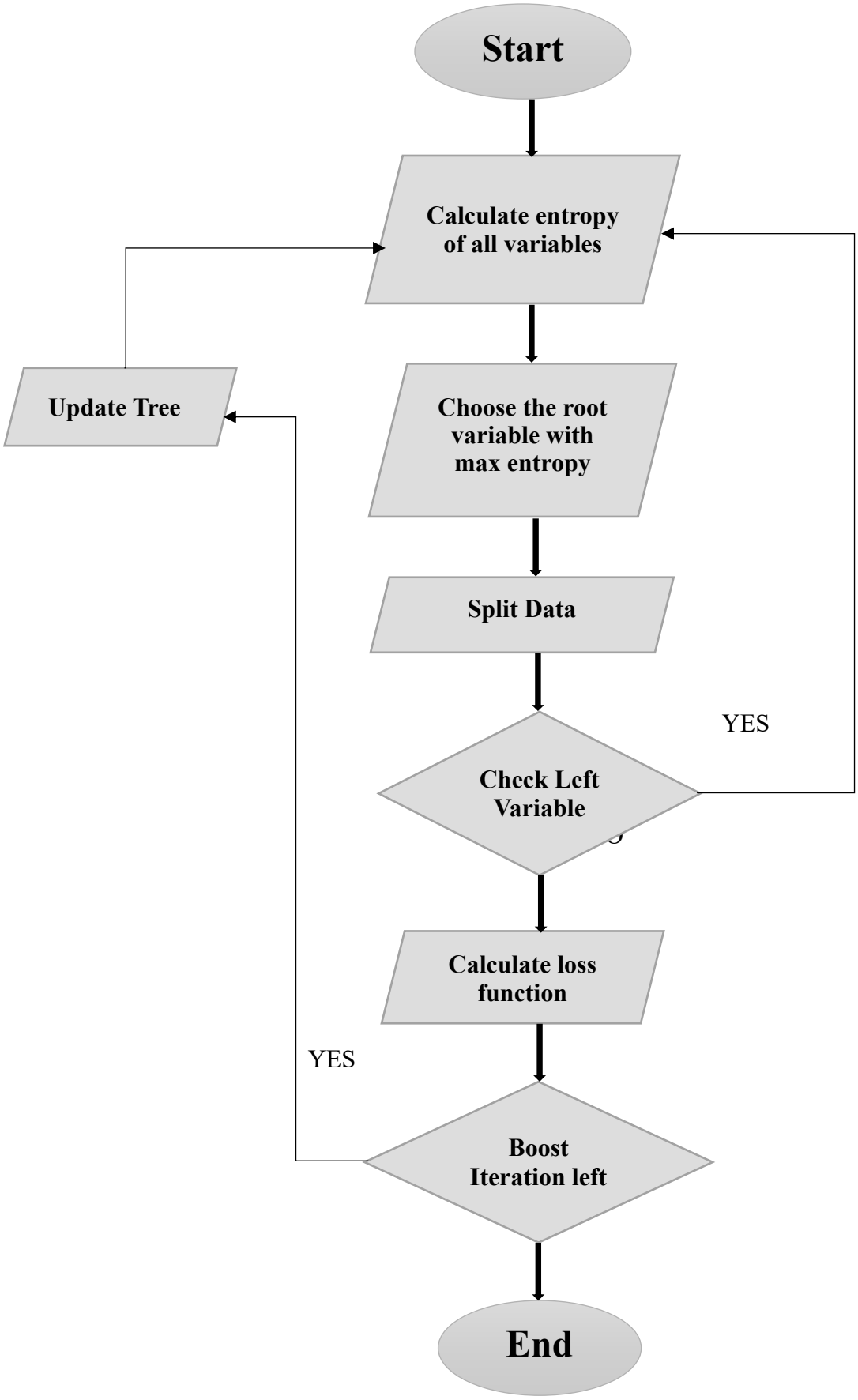
Step 5:

The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

Conclusion:

To sum up, after training several times our data(lung cancer data) using azure and code written from scratch, we achieved high accuracy from both training. we found a huge correlation between the people who are smoking, alcoholic and passive smoker. Based on our prediction passive smoker are most likely to get lung cancer more than even the people smoke by themselves. for instance, alcoholic and smoker have high chance of getting lung cancer. furthermore, the symptoms for getting lung cancer are fatigue, dry cough, chest pain and coughing blood.

Workflow


Start
Calculate entropy of all variables
Choose the root variable with max entropy
Split Data
Check Left Variable
YES
Calculate loss function
Boost Iteration left
YES
Update Tree
End

References

1. Eric (26 November 2017). Reasons to be optimistic about lung cancer, *The Star Malaysia*, Retrieved from https://www.pressreader.com
2. American Lung Association Scientific and Medical Editorial Review Panel (3rd November 2016). Lung cancer fact sheet, *American Lung Association.* Retrieved from https://www.lung.org
3. Kadir T., Gleeson F. (18th May 2015). Lung cancer prediction using machine learning and advanced imaging techniques, *Translational Lung Cancer Research,* Retrieved from https://www.ncbi.nlm.nih.gov
4. Siang K. C., MD1, John C. K. M. (1st June 2016). A Review of Lung Cancer Research in Malaysia, *Med J Malaysia,* Vol 70-71.
5. Leena P., Arpana S., Diksha K., Yogesh P. (February 2017). Lung Cancer Detection using Decision Tree Algorithm, *International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 02,* Page 1887-1888.
6. Joseph A. C., David S. W. (1st January 2006). Applications of Machine Learning in Cancer Prediction and Prognosis, *Sage Journals Cancer Informatics,* Retrieved from https://journals.sagepub.com
7. Neha P., Neha T., Surabhi B., Rewti A., Akshay G. (January 2015). A Survey on early detection and prediction of lung cancer, *International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 4,* pg.175 – 184

Figures

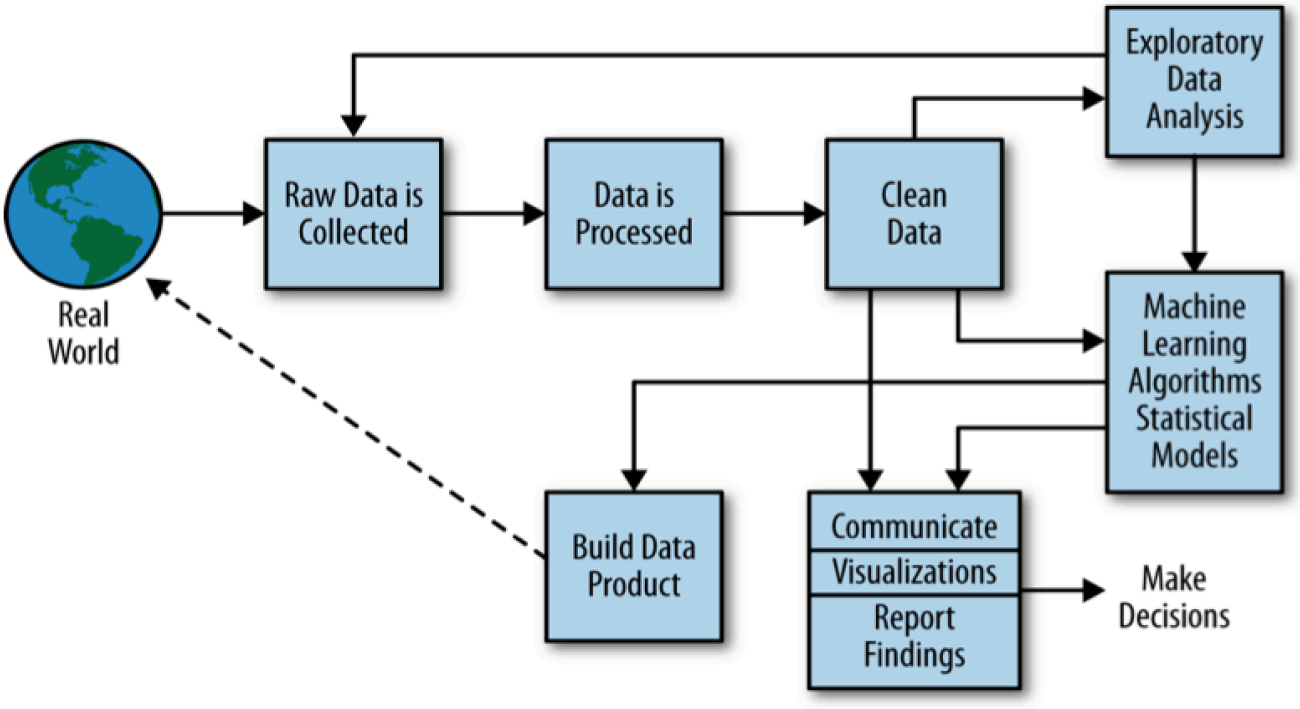


Figure 1: Data Science Process

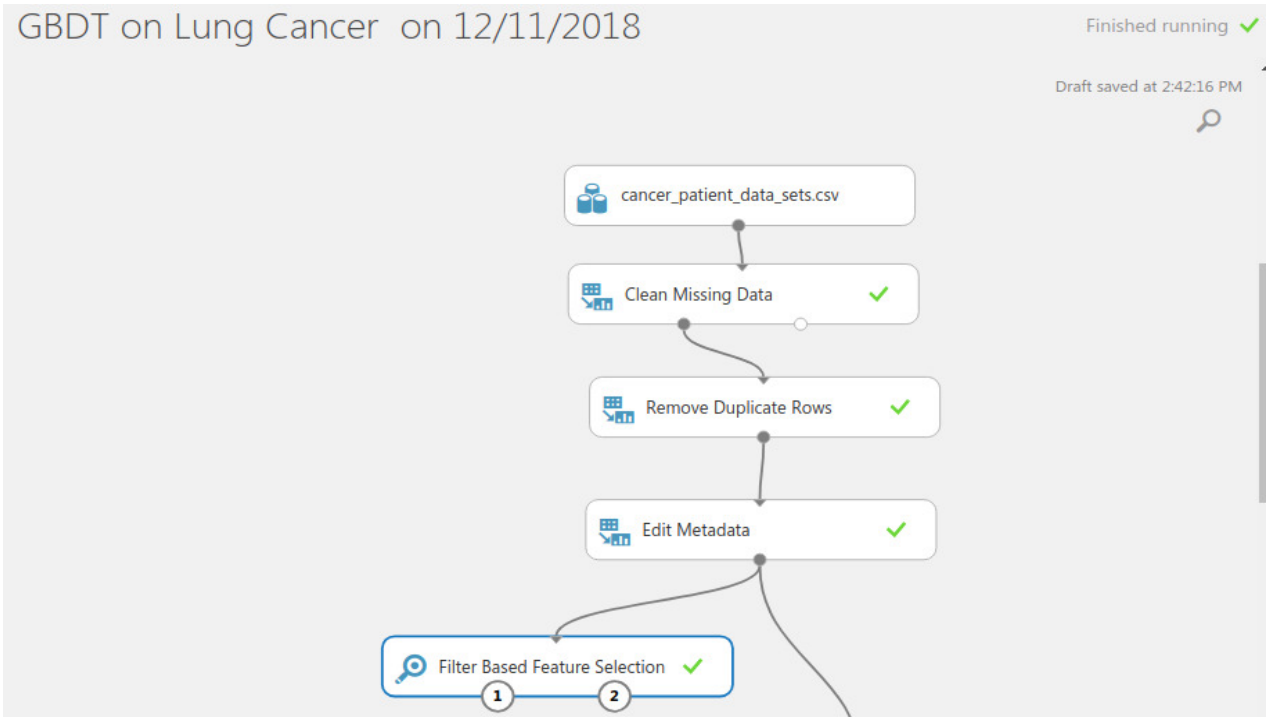


Figure 2: Data Processing

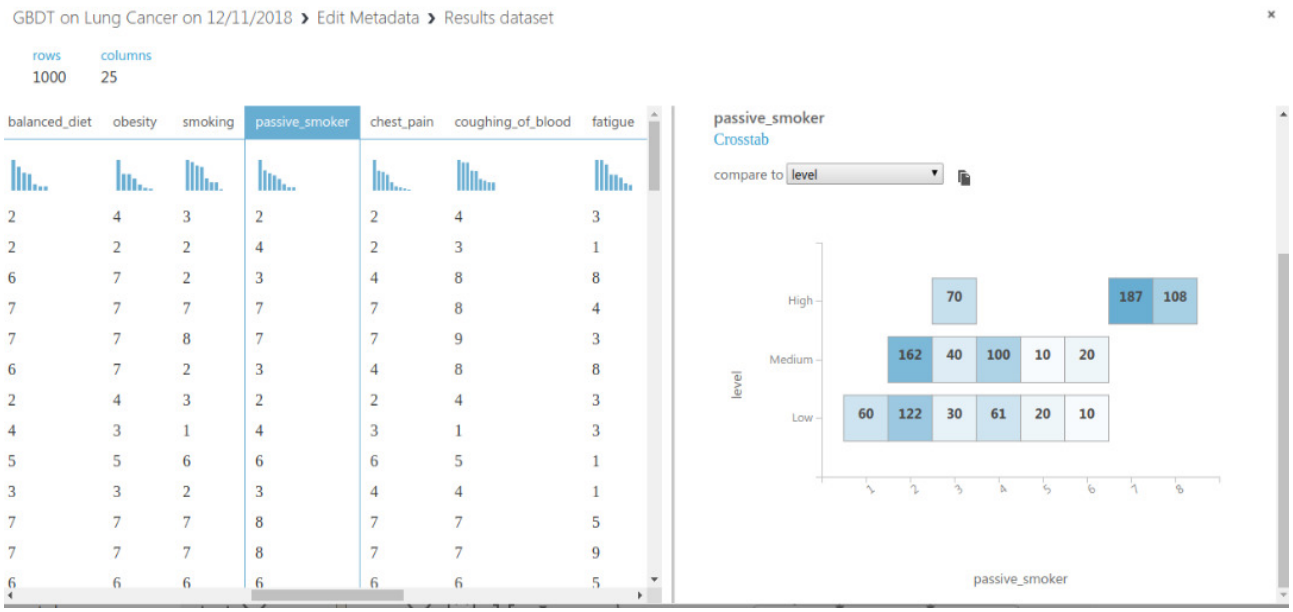


Figure 3a: Crosstab Analysis (Passive smoker vs level)

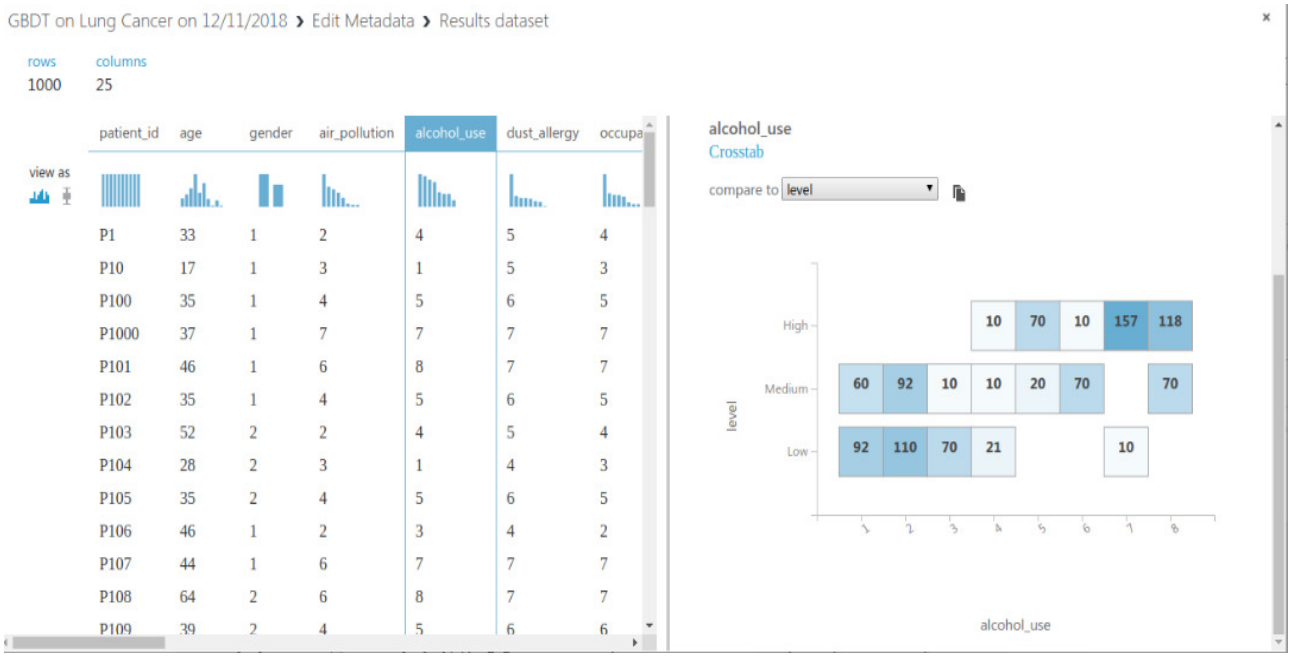


Figure 3b: Crosstab Analysis (alcohol vs level)

```
> roc_imp[order(roc_imp$Low,roc_imp$High,roc_imp$Medium),]
                              High       Low    Medium
age                       0.5666395 0.5804754 0.5804754
gender                    0.5993309 0.5993309 0.5508122
swallowing_difficulty     0.6754103 0.6754103 0.6625611
snoring                   0.7091912 0.7170315 0.7170315
dry_cough                 0.7340657 0.7340657 0.6486797
clubbing_of_finger_nails  0.7550839 0.7550839 0.7465953
weight_loss               0.7659885 0.7659885 0.7429023
frequent_cold             0.8214974 0.8214974 0.7092330
smoking                   0.8714309 0.8403635 0.8714309
shortness_of_breath       0.8604412 0.8604412 0.7667203
chronic_lung_disease      0.8922194 0.8922194 0.8189883
wheezing                  0.5780596 0.8995885 0.8995885
chest_pain                0.9091279 0.9091279 0.8469219
fatigue                   0.9135133 0.9135133 0.7555072
passive_smoker            0.9329084 0.9329084 0.9133520
air_pollution             0.9377910 0.9377910 0.8594653
dust_allergy              0.9431710 0.9431710 0.8427870
occupational_hazards      0.9447082 0.9447082 0.7397673
balanced_diet             0.9516705 0.9516705 0.9122380
genetic_risk              0.9561915 0.9561915 0.7629147
coughing_of_blood         0.9655500 0.9655500 0.9526737
obesity                   0.9805778 0.9805778 0.9078313
alcohol_use               0.9838148 0.9838148 0.7662981
```

Figure 4: Feature Importance