# **TOPIC: Building a patient diagnosis system using the KNN algorithm.**

Group ….

Bùi Đức Anh

Ngô Văn Chung

Nguyễn Văn Thành

Phạm Trung Hiếu

Lưu Văn Linh

**Contents**

[**TOPIC: Building a patient diagnosis system using the KNN algorithm.** 1](#_Toc90344620)

[**THANKS** 4](#_Toc90344621)

[**OPENING** 5](#_Toc90344622)

[**1.** **Reasons to choose the topic** 5](#_Toc90344623)

[**2.** **Urgency and purpose of the topic** 5](#_Toc90344624)

[**3.** **Objects, contents, methods and research range** 5](#_Toc90344625)

[**3.1.** **research objects** 5](#_Toc90344626)

[**3.2.** **research contents** 5](#_Toc90344627)

[**3.3.** **research methods** 5](#_Toc90344628)

[**3.4.** **research range** 6](#_Toc90344629)

[**4.** **The achievement** 6](#_Toc90344630)

[**CHAPTER I: BASIS THEORETICAL OF MACHINE LEARNING** 7](#_Toc90344631)

[**1.1.** **synopsis of Machine Learning** 7](#_Toc90344632)

[**1.2.** **Machine Learning in reality** 7](#_Toc90344633)

[**1.2.1.** **Image recognition** 7](#_Toc90344634)

[**1.2.3.** **Medical diagnosis** 8](#_Toc90344635)

[**1.3.** **How Machine Learning works** 8](#_Toc90344636)

[**1.4.** **Classification of Machine Learning Algorithms** 8](#_Toc90344637)

[**1.4.3.** **Supervised Learning** 9](#_Toc90344638)

[**1.4.4.** **Unsupervised Learning** 9](#_Toc90344639)

[**CHAPTER II: OVERVIEW OF K-NEAREST NEIGHBOR (KNN) ALGORITHM** 10](#_Toc90344640)

[**2.1. Synopsis of K-Nearest Neighbor Algorithm** 10](#_Toc90344641)

[**2.2. How does KNN work** 10](#_Toc90344642)

[**2.3. Application of K-Nearest Neighbor** 11](#_Toc90344643)

[**2.4. Advantages and disadvantages of K-Nearest Neighbor** 11](#_Toc90344644)

[**CHAPTER III: DESIGNING THE PATIENT'S DIAGNOSTIC SYSTEM USING KNN ALGORITHMS** 13](#_Toc90344645)

[**3.1. General model of the system** 13](#_Toc90344646)

[**3.2. Data preparation phase** 13](#_Toc90344647)

[**3.3. Stage of building the problem on Python** 15](#_Toc90344648)

[**3.4. Predicted results** 18](#_Toc90344649)

[**3.4.1. Results:** 18](#_Toc90344650)

[**3.4.2. Rate the accuracy:** 19](#_Toc90344651)

[**3.5. Design application interface** 20](#_Toc90344652)

[**3.5.1. Create the interface window:** 20](#_Toc90344653)

[**3.5.2. Create Textboxes and Edit textboxes to import the input:** 21](#_Toc90344654)

[**3.5.3. Create buttons to show the result:** 22](#_Toc90344655)

[**3.5.4. The finished interface:** 22](#_Toc90344656)

# **THANKS**

There is no student success without the enthusiastic and dedicated support of teachers, lecturers and professors. During the time since we started studying and making reports, we have received a lot of help from friends and teachers. We would like to express our sincere thanks to the teachers of the subject "Recognition Techniques", especially Ms. Nguyen Thi Thu for guiding and giving suggestions in each class session as well as during the time we were working on this report.

In the process of making this report, it is difficult to avoid mistakes, I hope you will forgive them. At the same time, due to limited qualifications and practical experience, shortcomings cannot be avoided. We look forward to receiving comments from teachers to accumulate experience for future products.

Finally, we would like to wish you and your teacher a lot of health, success in your career, and pass on knowledge to the next generation.

Hanoi, October 30, 2021

Student group made

# **OPENING**

## **Reasons to choose the topic**

When people around smoke, breathe in dust, polluted air, factory exhaust, getting respiratory diseases is inevitable, sooner or later. In all of respiratory diseases, we all know that lung cancer is the evilest and scariest disease. Therefore, our group of students, with the guidance of teachers, the help of friends and motivation from family, joined the project to create a program to diagnose lung cancer, using KNN algorithms. Our wish is to be able to contribute to the community and the country become healthier and prevent dangerous diseases.

## **Urgency and purpose of the topic**

In 2018, there were an estimated 9.6 million deaths from lung cancer. Lung cancer tops the list if one talks about their types and shares. Lung cancer cases are estimated to be around 2.09 million with 1.76 million deaths, accounting for about 84% of all deaths. For this reason, lung cancer has been considered as one of the deadliest diseases.

It is important that cancer is detected as early as possible because it tends to spread and cannot be cured in larger cases. It is difficult to diagnose lung cancer because it has symptoms in the late stages and it is almost impossible to save a person's life in the final stage.

To identify lung cancer, image processing and deep learning methods will be used. Accuracy can be improved using these methods. Detection and identification of tumors in terms of form, size and location is a difficult task. Timely detection saves a lot of time. And this time can be used to treat patients early.

## **Objects, contents, methods and research range**

### **research objects**

* General theory of identification techniques.
* KNN algorithm.

### **research contents**

* Basis theoretical of machine learning
* Overview of KNN algorithm
* Designing the patient's diagnostic system using KNN algorithms

### **research methods**

This report is based on a data collection method that combines both theory and experiment

* Collect research materials, including theoretical and experimental documents.
* Learn, survey techniques and technologies, available software, available on the market.
* Build an overview model of gold price identification
* Write code, test software.
* Compare with available software to make system performance evaluation

### **research range**

Books and real life experiment

## **The achievement**

# **CHAPTER I: BASIS THEORETICAL OF MACHINE LEARNING**

## **synopsis of Machine Learning**

Machine learning is a subfield of artificial intelligence. Instead of relying on explicit programming, it is a system through which computers use a massive set of data and apply algorithms to "train" on--to teach themselves--and make predictions.

Machine learning is a branch of AI. Other tools for reaching AI include rule-based engines, evolutionary algorithms, and Bayesian statistics. While many early AI programs, like IBM's Deep Blue, which defeated Garry Kasparov in chess in 1997, were rule-based and dependent on human programming, machine learning is a tool through which computers have the ability to teach themselves, and set their own rules. In 2016, Google's DeepMind beat the world champion in Go by using machine learning--training itself on a large data set of expert moves.

## **Machine Learning in reality**

### **Image recognition**

Image recognition is a well-known and widespread example of machine learning in the real world. It can identify an object as a digital image, based on the intensity of the pixels in black and white images or color images.

* Label an x-ray as cancerous or not
* Assign a name to a photographed face (aka “tagging” on social media)
* Recognize handwriting by segmenting a single letter into smaller images
  + 1. **Speech recognition**

Machine learning can translate speech into text. Certain software applications can convert live voice and recorded speech into a text file. The speech can be segmented by intensities on time-frequency bands as well.

Real-world examples of speech recognition:

* Voice search
* Voice dialing
* Appliance control

### **Medical diagnosis**

Machine learning can help with the diagnosis of diseases. Many physicians use chatbots with speech recognition capabilities to discern patterns in symptoms.

Real-world examples for medical diagnosis:

* Assisting in formulating a diagnosis or recommends a treatment option
* Oncology and pathology use machine learning to recognize cancerous tissue
* Analyze bodily fluids

ETC…

## **How Machine Learning works**

Machine learning uses two types of techniques: **supervised learning**, which trains a model on known input and output data so that it can predict future outputs, and **unsupervised learning**, which finds hidden patterns or intrinsic structures in input data.

## **Classification of Machine Learning Algorithms**

There are two common ways of grouping Machine learning algorithms. One is based on the learning style, the other is based on the function (of each algorithm).

* + 1. **Grouping based on learning method**
* Supervised Learning
* Classification
* Regression (Regression)
* Unsupervised Learning
  + Clustering (grouping)
  + Association
* Semi-Supervised Learning
* Reinforcement Learning
  + 1. **Function-based grouping**
* Regression Algorithms
* Classification Algorithms
* Instance-based Algorithms
* Regularization Algorithms
* Bayesian Algorithms
* Clustering Algorithms
* Artificial Neural Network Algorithms
* Dimensionality Reduction Algorithms
* Ensemble Algorithms

### **Supervised Learning**

[**Supervised machine learning**](https://www.mathworks.com/discovery/supervised-learning.html) builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Use supervised learning if you have known data for the output you are trying to predict.

### **Unsupervised Learning**

[**Unsupervised learning**](https://www.mathworks.com/discovery/unsupervised-learning.html) finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labeled responses.

In this project, we focus on the unsupervised learning technique.

# **CHAPTER II: OVERVIEW OF K-NEAREST NEIGHBOR (KNN) ALGORITHM**

## **2.1. Synopsis of K-Nearest Neighbor Algorithm**

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

## **2.2. How does KNN work**

The K-NN working can be explained on the basis of the below algorithm:

* **Step-1:** Select the number K of the neighbors
* **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
* **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
* **Step-4:** Among these k neighbors, count the number of the data points in each category.
* **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
* **Step-6:** Our model is ready.

## **2.3. Application of K-Nearest Neighbor**

K-nearest algorithm is used in various sectors of day to day life. It is easy to use so that data scientists and the beginner of machine learning use this algorithm for a simple task. Some of the uses of the k nearest neighbor algorithm are:

* **Finding diabetics ratio**

Diabetes diseases are based on age, health condition, family tradition, and food habits. But is a particular locality we can judge the ratio of diabetes based on the K Nearest Neighbor Algorithm. If you figure out the data of is age, pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index and other required data we can easily plot the probability of diabetes at a certain age.

* **Recommendation System**

If we search any product to any online store it will show the product. Decide that particular product it recommends some other product. You will be astonished after knowing that the 35% revenue of Amazon comes from the recommendation system. Decide the online store, Youtube, Netflix, and all search engines use the algorithms of k-nearest neighbor.

* **Concept Search**

Concept search is the industrial application of the K Nearest Neighbor Algorithm. It means searching for similar documents simultaneously. The data on the internet is increasing every single second. The main problem is extracting concepts from the large set of databases. K-nearest neighbor helps to find the concept from the simple approach.

* **Finding The Ratio of Breast Cancer**

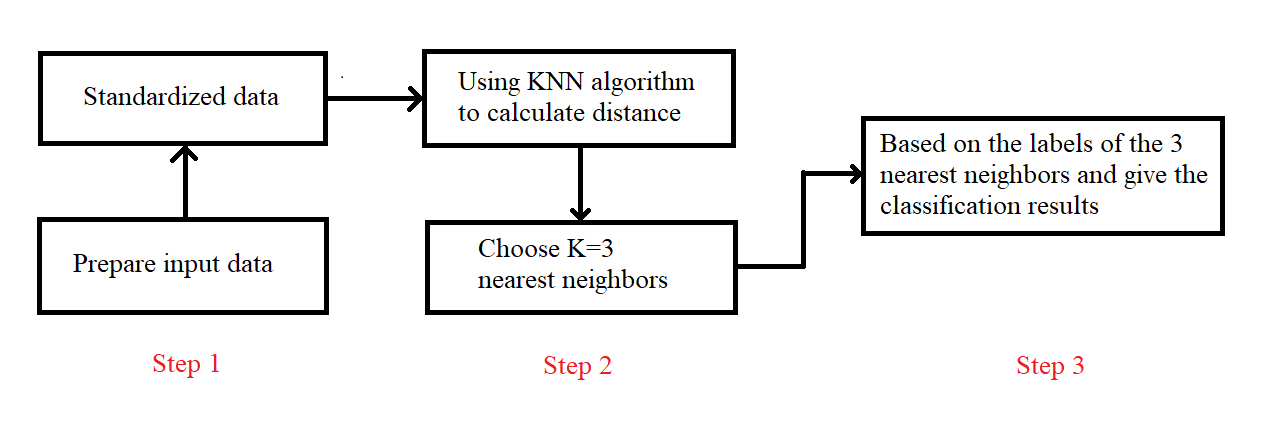
In the medical sector, the KNN algorithm is widely used. It is used to predict breast cancer. Here KNN algorithm is used as the classifier. The K nearest neighbor is the easiest algorithm to apply here. Based on the previous history of the locality, age and other conditions KNN is suitable for labeled data.

## **2.4. Advantages and disadvantages of K-Nearest Neighbor**

* ***Advantages of K-nearest Neighbor***
* It is simple to implement
* It is robust to the noisy training data
* It can be more effective if the training data is large
* ***Disadvantages of K-nearest Neighbor***
* Always need to determine the value of K which may be complex some time
* The computation cost is high because of calculating the distance between the data points for all the training samples

# **CHAPTER III: DESIGNING THE PATIENT'S DIAGNOSTIC SYSTEM USING KNN ALGORITHMS**

## **3.1. General model of the system**



*Figure 3.1: Functional block diagram of the system*

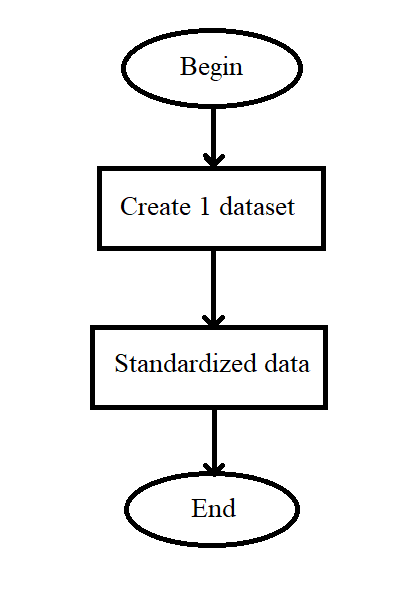
**Step 1:** Our input data is the customer's attributes designed by us. Then we will normalize the data so that the data attributes avoid mutual influence and create a premise to go to step 2.

**Step 2:** From the normalized properties from step 1, apply the KNN algorithm specifically, using the formula to calculate the Euclidean distance to find the distance between the Query Point and all the points in the data set.

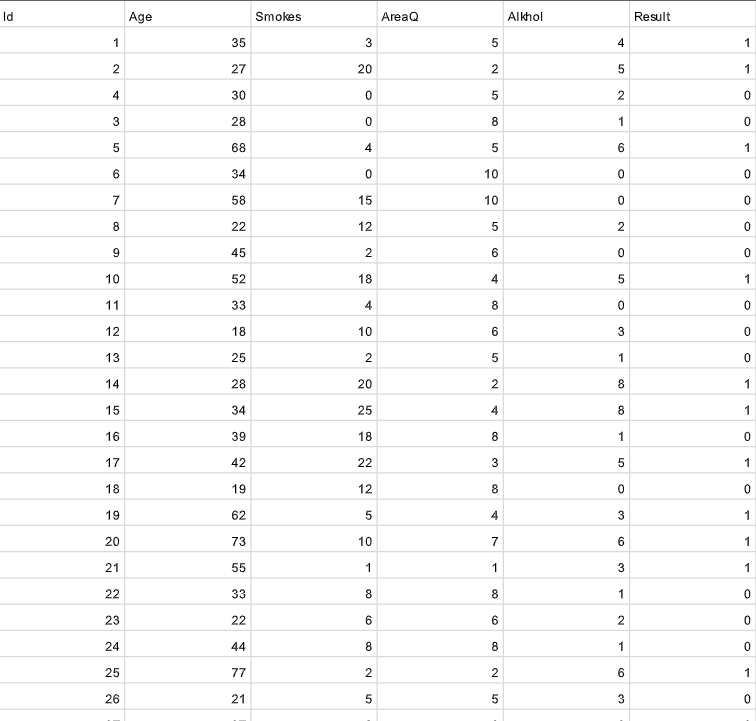
**Step 3:** Identify 3 nearest neighbors and based on the output of those 3 neighbors will give classification results.

## **3.2. Data preparation phase**

At this stage we have created our own dataset of 150 diseases.



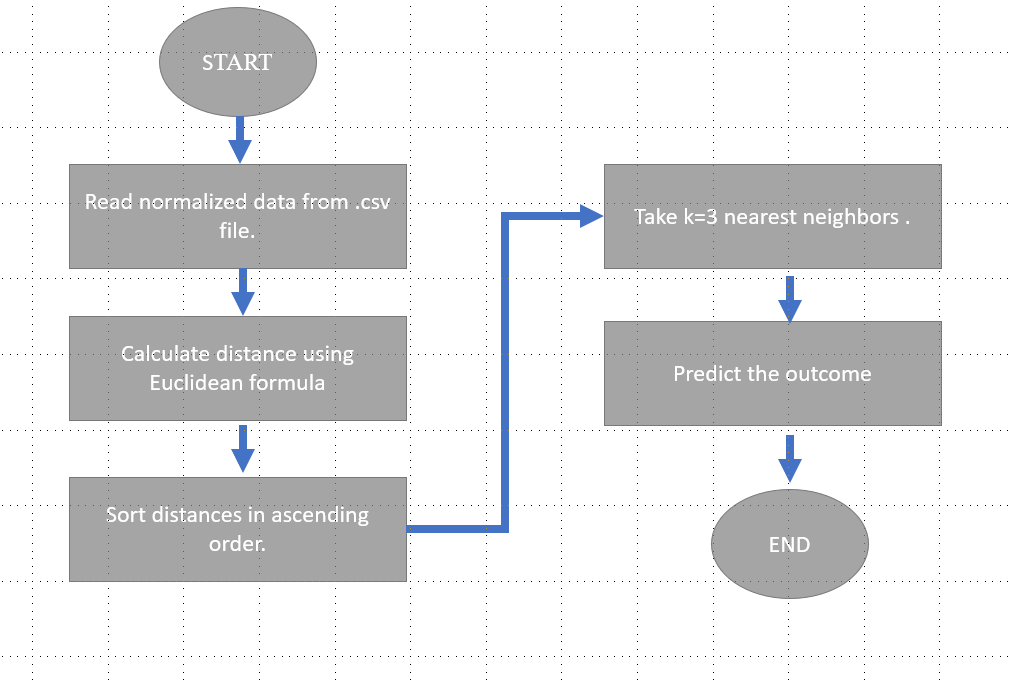
*Figure 3.2: Algorithm diagram of data preparation stage*

**

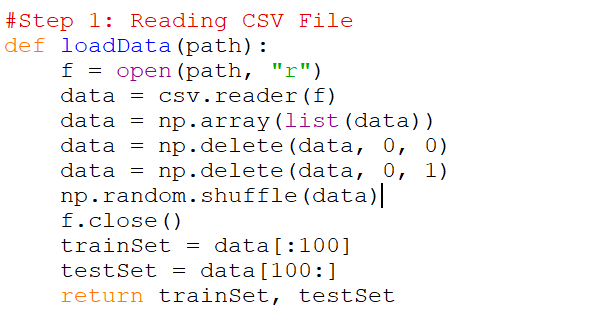
*Figure 3.3: Patient data set*

Then from this data set, we will use the min-max formula to normalize the numeric data from 0 to 1, so that the numeric data avoids affecting the alphanumeric data in the calculation process. distance.

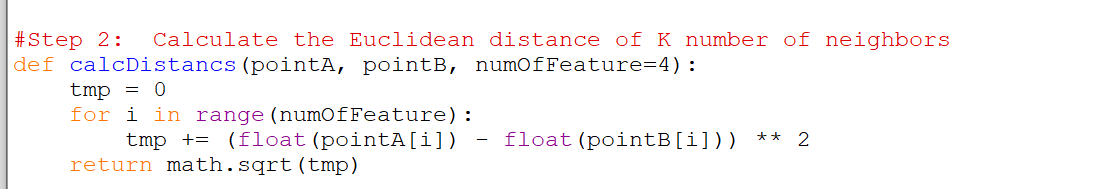
## **3.3. Stage of building the problem on Python**



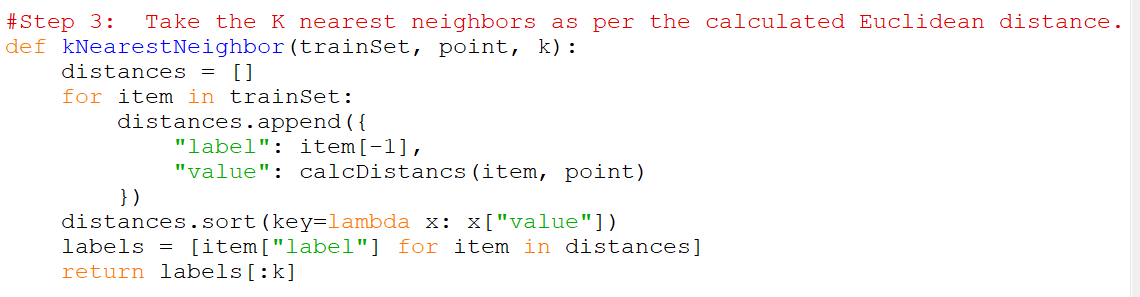
Step 1: **Read data at file .csv to get property**



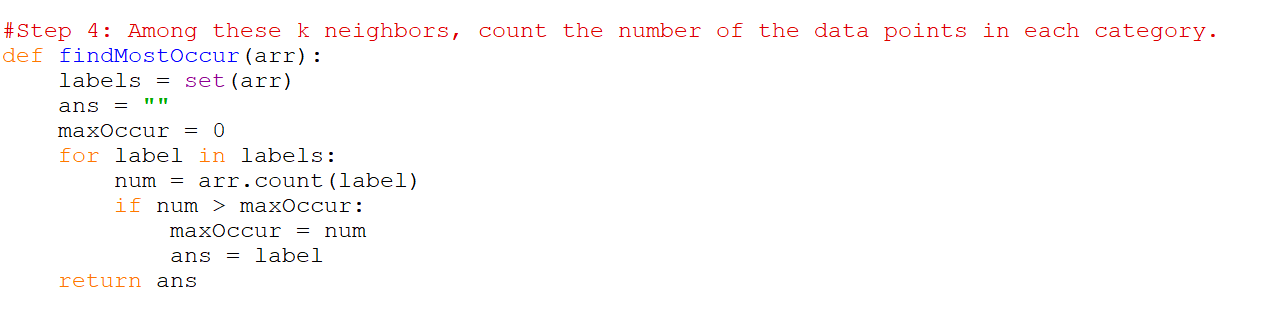
**Step 2: Calculate the Euclidean distance of k number of neighbors**



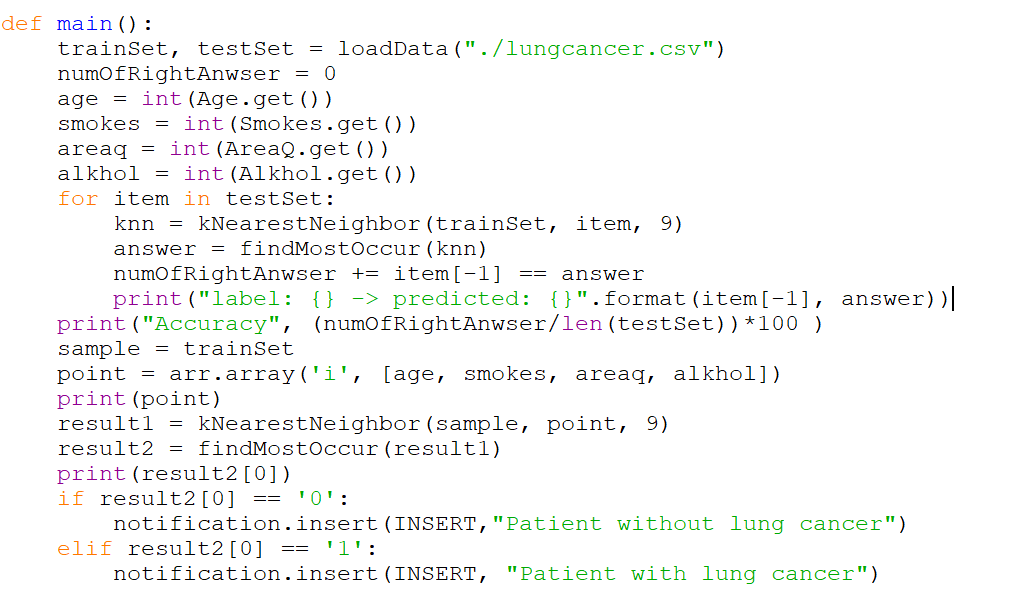
**Step 3: Take the k nearest neighbors as per the calculated Euclidean distance**



**Step 4: Among these k neighbors, count the number of the data points in each category.**



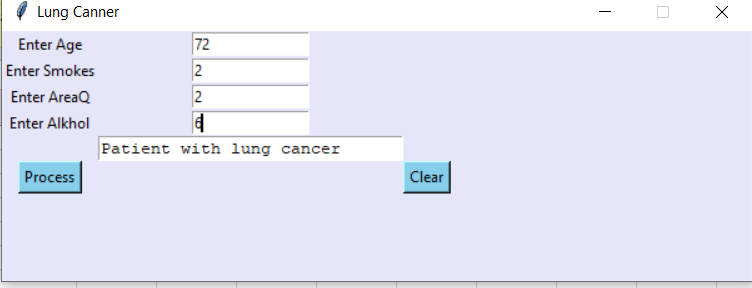
**Step5: Give the final result and the accuracy to end the program.**



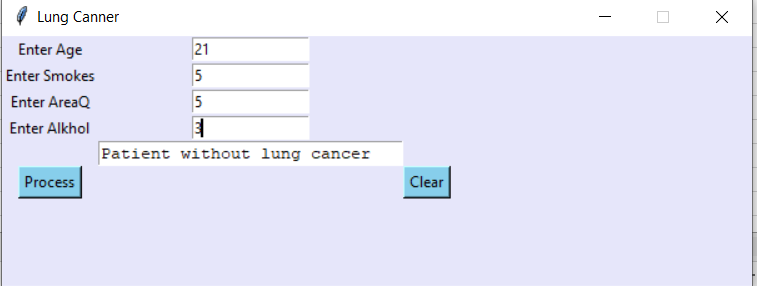
## **3.4. Predicted results**

### **3.4.1. Results:**

The patient with Age = 72, Smokes = 2, AreaQ = 2, Alkhol = 6 will have the result is “Patient with lung cancer”.

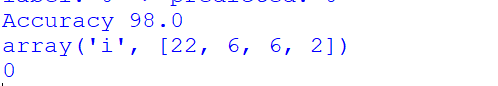


The patient with Age = 21, Smokes = 5, AreaQ = 5, Alkhol = 3 will have the result is “Patient without lung cancer”.



### **3.4.2. Rate the accuracy:**

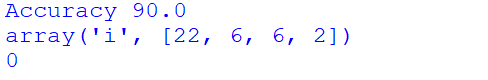
With k=3, the calculated accuracy is 98%.



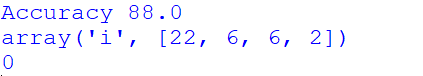
With k=5, the calculated accuracy is 92%.



With k=7, the calculated accuracy is 90%.



With k=9, the calculated accuracy is 88%.



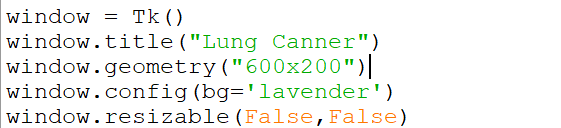
We found that when K=3 nearest neighbor, we’ll get the highest accuracy at the result. Other else, if we increase K, the accuracy will have a light fluctuate but not affect too much the diagnosis result. So, in this topic, we choose K = 3 nearest neighbor for the best result.

If we divide different ratios, we’ll get different results, the larger dataset, the more accurate the classification result.

## **3.5. Design application interface**

### **3.5.1. Create the interface window:**

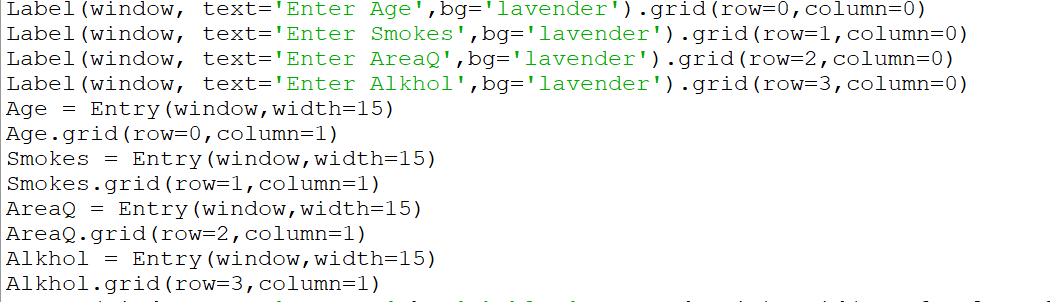
Create a window called Tk has the definition 600x200, “Lavender” color, title it “Lung Cancer”.



After finish creating the window above, this is the result:



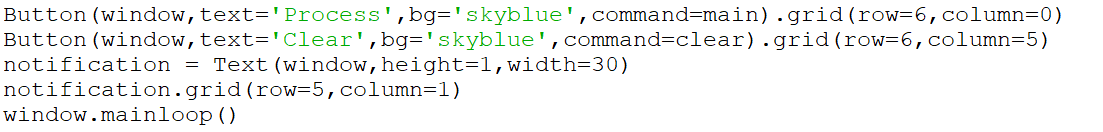
### **3.5.2. Create Textboxes and Edit textboxes to import the input:**



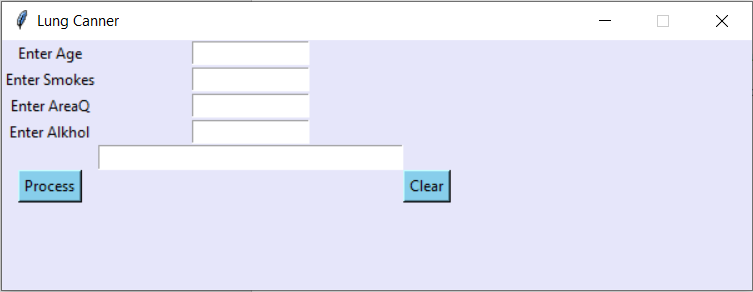
**Result:**



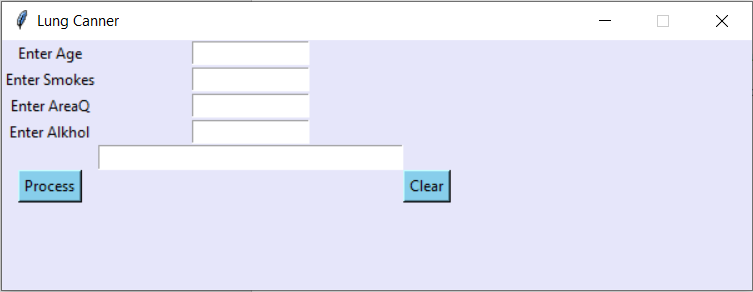
### **3.5.3. Create buttons to show the result:**



**Result:**



### **3.5.4. The finished interface:**



**CONCLUSIONS AND DEVELOPMENT DIRECTIONS**

**DOCUMENTS FOR REFERENCE**