LUNG CANCER PREDICTION

An HealthCare Training Project Report

Submitted to faculty of

JAWAHARLAL NEHRU TECHNOLOGY UNIVERSITY , HYDERABAD

HYDERABAD

In partial fulfilment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

by

ANSUMAN PASWAN MANGALARAPU SHIRISHA

(177Y1A0528),CSE Dept (177Y1A0548),CSE Dept

DEVARAKONDA SREESHMA TADAKA RAHUL

(177Y1A0450),ECE Dept (187Y1A0492),ECE Dept



MARRI LAXMAN REDDY INSTITUTE of TECHNOLOGY

and MANAGEMENT

(Approved by AICTE, New Delhi & Affilated to JNTU,HYDERABAD)

DUNDIGAL VILLAGE ,

MEDCHAL: 500043

TELANGANA

MARRI LAXMAN REDDY INSTITUTE of TECHNOLOGY

and MANAGEMENT

(Approved by AICTE, New Delhi & Affilated to JNTU, HYDERABAD)



CERTIFICATE

This is certify that the Health care Project Report entitled “**Lung Cancer Prediction**” is a Bonafide record of work carried Out by ANSUMAN.P , SHIRISHA.M , SREESHMA.D and RAHUL.T in partial fulfilment of the requirements for the award of the Degree of Bachelor Technology and Engineering of Jawaharlal Nehru TechnologyUniversity Hyderabad Academic Year- 2017-2021

Head of the Department

Health Care Training Coordinator



INTERNSHIP REPORT APPROVAL FORM

June 20,2019

With immense pleasure ,this is to approve that the students of Marri Laxman Reddy Institute of technology and Management

**Ansuman Paswan(177y1a0528)**

**Shirisha(177y1a0548)**

**Sreeshma(177y1a0450)**

**Rahul(187y1a0492)** Succesfully completed their Project and Project Report on “**Lung Cancer** **Prediction** “Using **Random Forest** under our guidance

We are highly impressed with the work that they have done and commened them on their quik grasping skills .They have shown good interest to learn and have put the knowledge gained into application in the form of this project .We appreciate the hard work and the commitment to learn things shown by them .

We, hereby approve that this document is completely checked and accepted by SmartBridge Technical Team its been an absolute pleasure to Educate and Mentor these Students. We hope that this document will also serve as a Letter of Recommendation ,to whomsover applied

We wish them success in all future endevors and a great career ahead

Anil Chowdhary

AI & ML Developer

**ACKNOWLEDGEMENT**

The Satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragement crown all the efforts with success.

We would like to take the opportunity to thank our beloved principal **Dr. K. Venkateswara Reddy** for providing a great support for us in completing our projects and giving us opportunity to be a part of the Internship Program

We feel elated to express our floral gratitude and sincere thanks to **Dr.K.Abdul Bashith ,** Head of the Department, Computer Science and Engineering for his encouragements all the way during analysis of project His annotations, encouragement and criticism are the key behind the successful completion of project work.

We feel elated to express our floral gratitude and scincere thanks to **B. Prasad** Associate Professor, Computer Science and Engineering Department the person who made these event possible . who encouraged us and made to know the importance of the Program.

We would like to take the opportunity to thank **Smart Bridge Educational** **Service Pvt . Ltd** for providing this training program and Support for us completing the project. Our team would like to thank **Ramya** Mam (Trainer)for making us understand the concept so well. Next we would like to thank our Mentor **Gokul** Sir for clarifying our doubts and helping us each and every step during this Program.

Our special thanks to the faculty of our Department and Programmers of our Computer lab. Finally, we would like to thank our family members ,non-teaching staff ,attendants and our friends ,who directly or indirectly helped and supported us in completing our project in time.

**ABSTRACT**

It has been witnessed that the Person who Drinks and Smokes regulary have a great chance to get infected to Disease **Lung Cancer .**Now a days this disease have been very common due to attraction towards the Alcohol and Cigrette at very young age .The main obejective of **lung Cancer Prediction** is predict the chance to person can be affected by the disease .By giving the input as the Age of the person ,smokes per day and the amount of alchohol the person drinks .by taking the above inputs we can predict the person can get affected or not with the disease. In this we are going to use Dataset from which we are going to train our model .By giving the input and predicting our result.by using the appropriate regression mlodel.

INDEX

TITLE PAGE NO

**MOTIVATION** 1

**SCOPE** 2

**ABSTRACT** 3

CHAPTER 1:

1.1: Introduction 4-7

1.2: Objective of Research 8

1.3: Problem Statement 9

CHAPTER 2: **REVIEW OF LITERATURE** 10-11

CHAPTER 3: **DATA COLLECTION** 12

CHAPTER 4: **METHODOLOGY**

4.1: Expolatary Model 13

4.2: Modeling 14-15

CHAPTER 5: **REFERENCE** 16

CHAPTER 6: **CONCLUSION** 17

**MOTIVATION:**

In today’s world many persons are getting affected to the disesse Cancer there are several kind of cancer Lung Cancer is one of them which is a life taking disease. Which is normally caused due to Smoking and Drinking Alcohol. and get aware of the disease.

**PROBLEM:**

Recognising the person who is affected with the disease is not a easy task to do .it need a proper doctor checkup before checkup they may not know from what they are suffering.

**SOLUTION:**

With the help of the previous Dataset we will teach our Model for predicting the disease .By giving the inputs of previous affected person from the disease when we’ll give the inputs the model will tell is the person is going to be affected by the disease or not.

**SCOPE**

Used for predicting the person is going to affected by the disease by analysing the habits of the person. Like drinking, smoking etc. with the help of the inputs from the trained model we will predict the result.

**PROJECT REPORT**

**LUNG CANCER PREDICTION USING RANDOM FOREST**

* 1. **Introduction**

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aims to help programmers write clear, logical code for small and large-scale projects

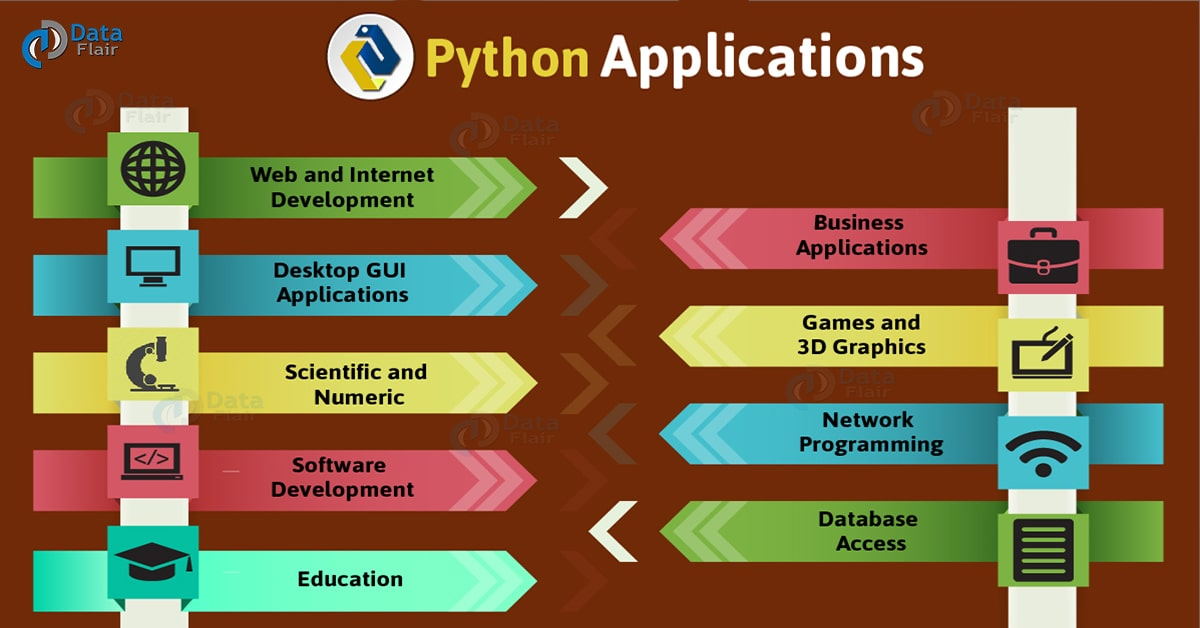
Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural , object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library

Python was conceived in the late 1980s as a successor to the ABC language. Python 2.0, released 2000, introduced features like list comprehensions and a garbage collection system capable of collecting reference cycles. Python 3.0, released 2008, was a major revision of the language that is not completely backward-compatible, and much Python 2 code does not run unmodified on Python 3. Due to concern about the amount of code written for Python 2, support for Python 2.7 (the last release in the 2.x series) was extended to 2020.

Language developer Guido van Rossum shouldered sole responsibility for the project until July 2018 but now shares his leadership as a member of a five-person steering council**.**

Python interpreters are available for many operating systems. A global community of programmers develops and maintains CPython, an open source reference implementation.

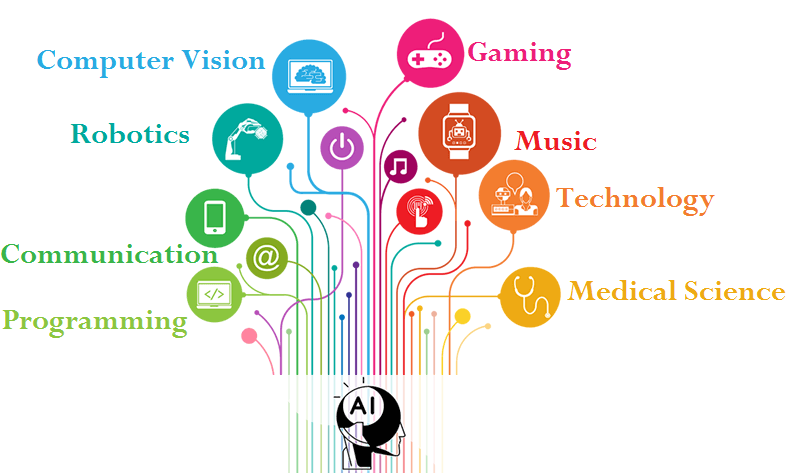
A non-profit organization, the Python Software Foundation, manages and directs resources for Python and C Python development.



**Fig : Application of Python**

**ARTIFICIAL INTELLIGENCE**

In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans. Colloquially, the term "artificial intelligence" is used to describe machines/computers that mimic "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving"



**Fig : AI Applications**

**Introduction to Lung Cancer:**

Lung cancer is the one of the leading cause of cancer deaths in human men. Manifestation of Lung cancer in the body of the patient reveals through early symptoms in most of the cases. Treatment and prognosis depend on the histological type of cancer, the stage (degree of spread), and the patient's performance

Possible treatments include surgery, chemotherapy, and radiotherapy Surviva l depends on stage, overall health, and other factors, but overall only 14% of people diagnosed with lung cancer survive five years after the diagnosis. Symptoms that may suggest lung cancer include

* dyspnea (shortness of breath with activity),
* hemoptysis (coughing up blood),
* chronic coughing or change in regular coughing pattern,
* wheezing,
* chest pain or pain in the abdomen,
* cachexia (weight loss, fatigue, and loss of appetite),
* dysphonia (hoarse voice),
* clubbing of the fingernails(uncommon),
* dysphasia(difficulty swallowing),
* Pain in shoulder ,chest , arm,
* Bronchitis or pneumonia,
* Decline in Health and unexplained weight loss.

**1.2. Objective of Research**

Prediction of survival in patients diagnosed with lung cancer remains problematical. The aim of the present study was to examine the clinical utility of an established objective marker of the systemic inflam matory response, the Glasgow Prognostic Score, as the basis of risk stratification in patients with lung cancer. Methods. Between 2005 and 2008 all newly diagnosed lung cancer patients coming through th e multidisciplinary meetings (MDTs) of four Scottish centres were included in the study. The details of 882 patients with a confirmed new diagnosis of any subtype or stage of lung cancer were collected prospectively. Results. The median survival was 5.6 months (IQR 4.8-6.5) . Survival analysis was undertaken in three separate groups based on mGPS score In the mGPS 0 group the most highly predictive factors were performance status, weight loss, stage of NSCLC, and palliative treatment offered. In the mGPS 1 group performance status, stage of NSCLC, and radical treatment offered were significant. In the mGPS 2 group only performance status and weight loss were statistically significant.

Discussion. This present study confirms previous work supporting the use of mGPS in predicting cancer survival; however, it goes further by showing how it might be used to provide more objective risk stratification in patients diagnosed with lung cancer.

1.2 **Problem Statement**

Lung cancer is the most common cancer among human in world Every thirteen minutes a woman dies with the diagnosis of Lung cancer. These facts have led researchers to continue studying how to treat and detect Lung cancer in human especially older human, who are of higher risk. Sonography (also known as ultrasound) has become a great addition to mammography and magnetic resonance imaging (MRI) as imaging techniques dedicated to providing Lung cancer screening.

**REVIEW OF LITERATURE**

The approach that is being followed here for the prediction technique is based on systematic study of the statistical factors, symptoms and risk factors associated with Lung cancer. Non-clinical symptoms and risk factors are some of the generic indicators of the cancer diseases.Initially the parameters for the pre-diagnosis are collected by interacting with the pathological, clinical and medical oncologists (Domain experts).

**A. Statistical Incidence Factors:**

i. Age-adjusted rate (ARR)

ii. Primary histology

iii. Area-related incidence chance

iv. Crude incidence rate

**B. Lung cancer symptoms:**

The following are the generic lung cancer symptoms

i. A cough that does not go away and gets worse over time

ii. Coughing up blood (heamoptysis) or bloody mucus.

iii. Chest, shoulder, or back pain that doesn't go away and often is made worse by deep Hoarseness

iv. Weight loss and loss of appetite

v. Increase in volume of sputum

vi. Wheezing

vii. Shortness of breath

Viii.Repeated respiratory infections, such as bronchitis or pneumonia

Ix. Repeated problems with pneumonia or bronchitis

**C. Lung cancer risk factors:**

**a. Smoking:**

i. Beedi

ii. Cigarette

iii. Hukka

b. Second-hand smoke

c. High dose of ionizing radiation

d. Radon exposure

e. Occupational exposure to mustard gas chloromethyl ether, inorganic arsenic, chromium, nickel, vinyl

chloride, radon asbestos

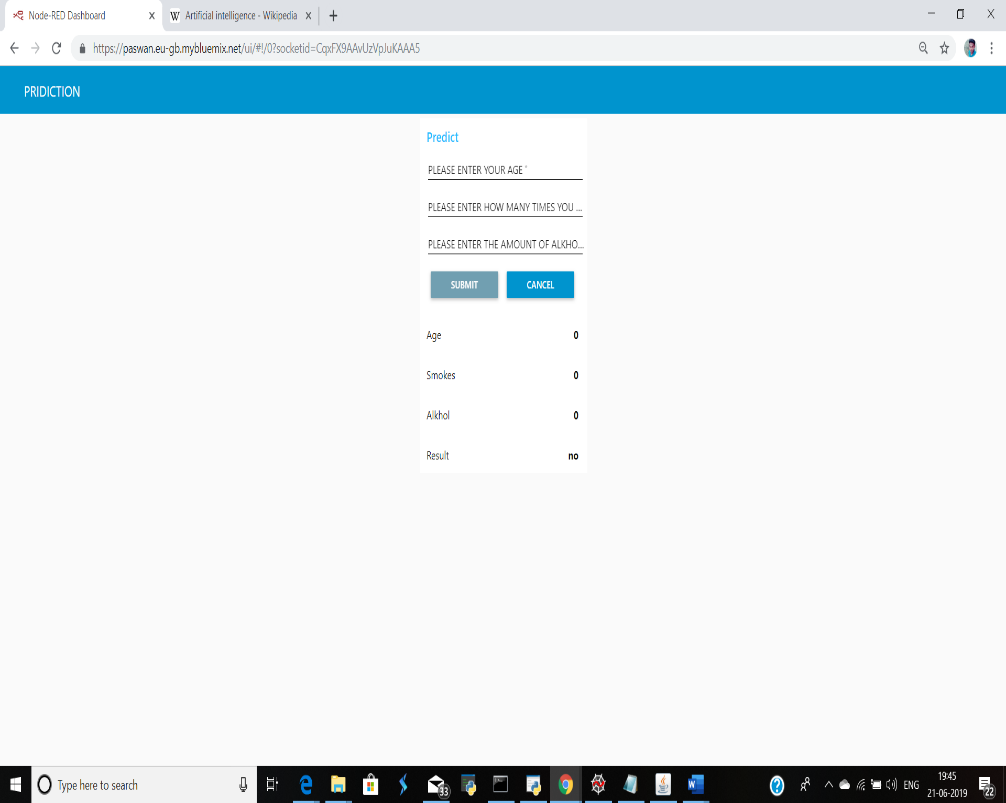
f. Air pollution

g. Insufficient consumption of fruits & vegetables

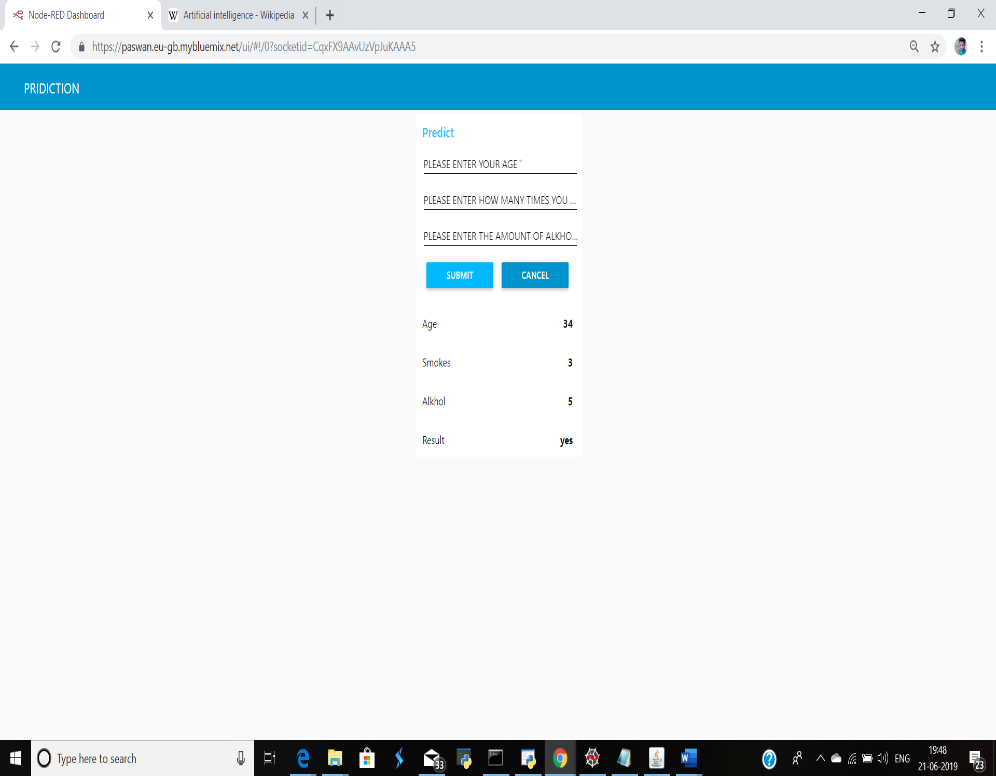
h. Suffering with other types of malignancy

**METHODOLOGY:**

**4.1** EXPLOTORY DATA ANALYSIS



**Fig : UI model of lung cancer prediction**

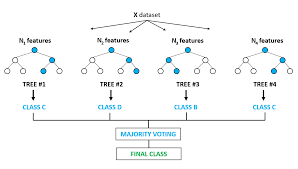


**Fig : A Sample Prediction**

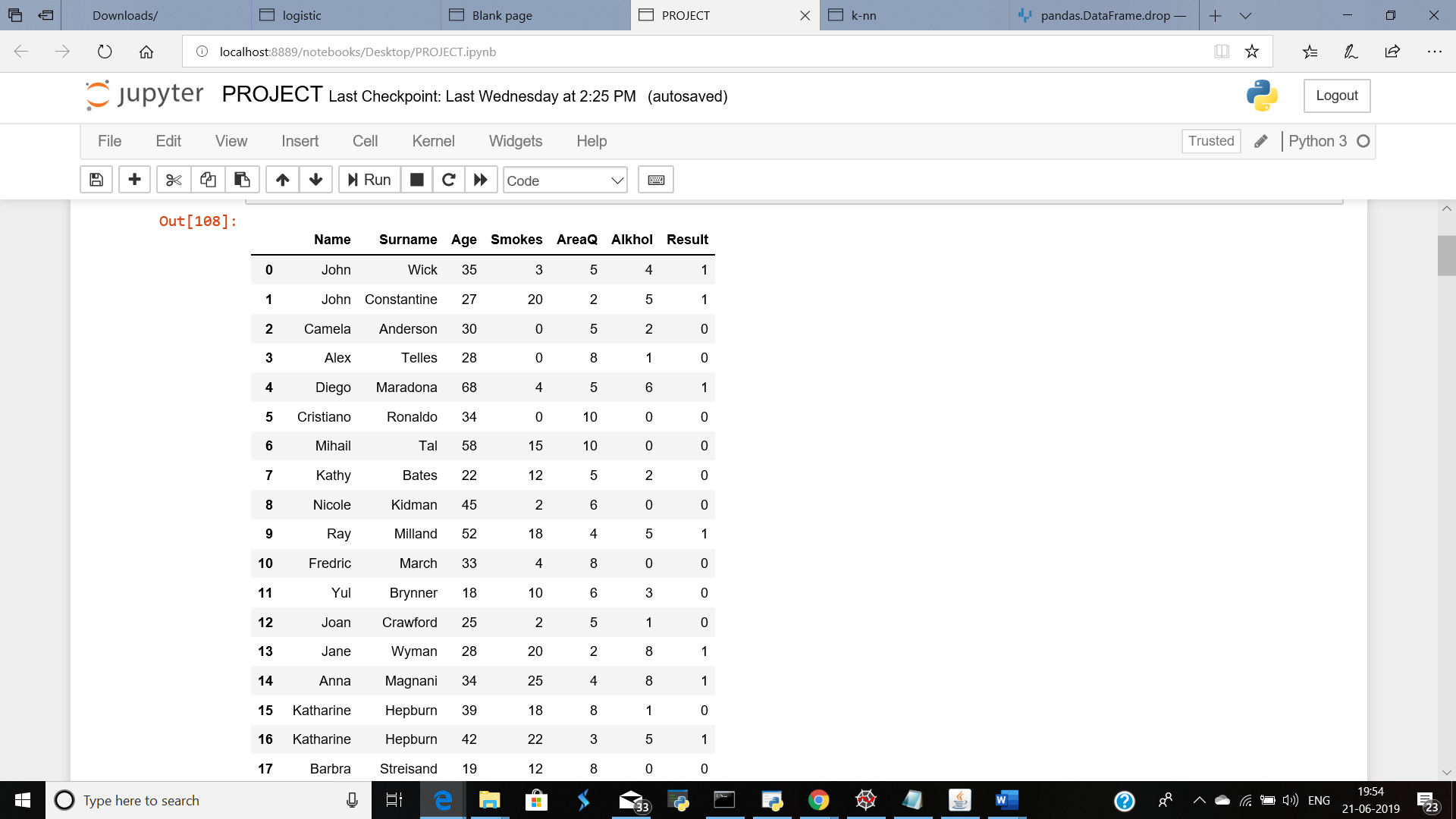
**4.2 DATA MODELING**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

The first algorithm for random decision forests was created by Tin Kam Housing the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg

****

**Fig: Random Forest Model**



**FIG: The data set used in the algorithm**

**6. CONCLUSION**

A prototype lung cancer disease prediction system is developed using data mining classification techniques. The system extracts hidden knowledge from a historical lung cancer disease database. The most effective model to predict patients with Lung cancer disease appears to be **Random Forest** . Random Forest results are easier to read and interpret. The drill through feature to access detailed patients’ profiles is only available in Random Forest. Naïve Bayes fared better than Random Forest as it could identify all the significant medical predictors. The relationship between attributes produced by Neural Network is more difficult to understand. In some cases even in the advanced level Lung cancer patients does not show the symptoms associated with the Lung cancer. Prevalence of Lung cancer disease is high in India, especially in rural India, did not get noticed at the early stage, because of the lack of awareness. Also it is not possible for the voluntary agencies to carry out the screening for all the people. The emphasis of this work is to find the target group of people who needs further screening for Lung cancer disease, so that the prevalence and mortality rate could be brought down. Lung cancer prediction system can be further enhanced and expanded. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining