**SURVIVAL ANALYSIS-CANCER PREDICTION**

**PROJECT REPORT**

**Submitted by**

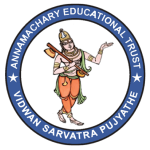
**CODE BLACK**

**N.SUPRIYA,T.SANJANAPRIYA,P.YASMIN,G.MADHUMITHA**

***In partial fulfilment for the award of the Certificate***

**of**

**SUMMER INTERNSHIP PROGRAM**

**Department of Computer Science and Engineering**

**Annamacharya Institute of Technology and Sciences**

**Venkatapuram Village , Renigunta Mandal , Tirupati , Andhra Pradesh 517520**

**July 2019.**

### BONAFIDE CERTIFICATE

This is to certify that the project entitled ”SURVIVAL ANALYSIS-CANCER PREDICTION”submitted by **N.SUPRIYA, T.SANJANAPRIYA ,P.YASMIN, G.MADHUMITHA** in partial fulfilment for the requirements for the award of internship certification in technologies of Machine learning and Deep learning is an authentic work carried out by them under my supervision and guidance.

To the best of my knowledge, the matter embodied in the project report has not been submitted to any other University/Institute for the award of any Degree or Diploma.

### Signature of Supervisor                                       Signature of Head of the Department

 Akshay Kumar Kothuri, M.Tech       Mrs.B.Rupa Devi M.Tech,Ph.D.,

AI & IoT Developer Assistant Professor & HOD,

SmartBridge Educational Services PVT LTD Department of CSE,

Hyderabad – 500096 AITS,Tirupati.

**1.SURVIVAL ANALYSIS-CANCER PREDICTION**

**(LUNG CANCER)**

* 1. **Introduction:**

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data. In this article, we’ll see basics of Machine Learning, and implementation of a simple machine learning algorithm using python.

**Lung cancer**, also known as **lung carcinoma** is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung. This growth can spread beyond the lung by the process of metastasis into nearby tissue or other parts of the body. Most cancers that start in the lung, known as primary lung cancers, are carcinomas. The two main types are small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC). The most common symptoms are coughing (including coughing up blood), weight loss, shortness of breath, and chest pains.

The vast majority (85%) of cases of lung cancer are due to long-term tobacco smoking. About 10–15% of cases occur in people who have never smoked. These cases are often caused by a combination of genetic factors and exposure to radon gas, asbestos, second-hand smoke, or other forms of air pollution. Lung cancer may be seen on chest radiographs and computed tomography (CT) scans. The diagnosis is confirmed by biopsy which is usually performed by bronchoscopy or CT-guidance.

Avoidance of risk factors, including smoking and air pollution, is the primary method of prevention. Treatment and long-term outcomes depend on the type of cancer, the stage (degree of spread), and the person's overall health. Most cases are not curable. Common treatments include surgery, chemotherapy, and radiotherapy.NSCLC is sometimes treated with surgery, whereas SCLC usually responds better to chemotherapy and radiotherapy.

Worldwide in 2012, lung cancer occurred in 1.8 million people and resulted in 1.6 million deaths. This makes it the most common cause of cancer-related death in men and second most common in women after breast cancer. The most common age at diagnosis is 70 years. Overall, 17.4% of people in the United States diagnosed with lung cancer survive five years after the diagnosis, while outcomes on average are worse in the developing world.

**1.2 Objective of research:**

This research investigates the survival status of the patient who had undergone surgery for lung cancer. Particularly it seeks independent variable patterns to determine the survival times and identifies the correlation among them. The dataset is collected from an online website named as kaggle.com based on haberman's survival dataset .The outcome of this project gives the predicted information about survival status of a lung cancer victim who had undergone surgery.

**1.3 Problem Statement:**

Lung cancer is a major public health problem. It is a significant cause of mortality and morbidity and is a national target area in the Government’s Health Strategy .Our project mainly focuses on survival status of lung cancer victims who had undergone surgery. A threshold range of 5years after surgery was taken in order to predict whether a patient survived 5(or more than) years or died within 5years.

**2. REVIEW OF LITERATURE**

Lung Cancer is one of the group of diseases characterized by the uncontrolled growth and spread of abnormal cells. If the spread is not controlled, it can result in death. Although the reason for many cancers, particularly those that occur during childhood, remains unknown, established cancer causes include lifestyle (external) factors, such as excess body weight, and non-modifiable (internal) factors, such as inherited genetic mutations, hormones, and immune conditions. These risk factors may act simultaneously or in sequence to initiate and/or promote cancer growth. Ten or more years often pass between exposure to external factors and detectable cancer.

While there has been much research regarding risk factors for lung cancer, there is a need to get an analysis about the lung cancer victim's survival status after they had undergone through surgery. The outcome of our project is nothing but a prediction about survival status of patients who undergone from surgery.

Any data analysis task or for performing operation requires good domain knowledge that helps to relate the data features and also can give accurate conclusion. On considering this, our project is designed in a way that keenly elevates the features of our data set and how it affects other feature.



Fig: Lungs before and after affected due to smoke and alcohol

**3. DATA COLLECTION**

Haberman’s data set contains data from the study conducted in University of Chicago’s Billings Hospital between year 1958 to 1970 for the patients who undergone surgery of lung cancer.

**3.1. Source**

<https://www.kaggle.com/yusufdede/lung-cancer-dataset#lung_cancer_examples.csv>

There are 7 attribute in this data set out of which 6 are features and 1 class attribute as below. Also, there are 59 instances of data.

1. Name

2. Surname

3. Age

4. Smokes

5.AreaQ

6.Alkohol

7.Result(Survival status)

**3.2.Name,Surname:**

They represent the name and surname of the patient who was going for surgery.

**3.3.Age:**

It represents the age of the patient.

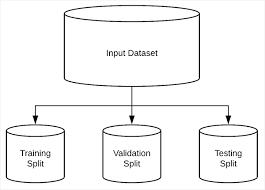
**3.4. Smokes:**

smoking is a strong risk factor for lung cancer. The risk of developing lung cancer is directly related to the number of “[pack-years](https://www.verywell.com/definition-of-pack-years-of-smoking-2249140)" a person smoked.

**3.4.Result( Survival Status):**

It represent whether patient survive more than 5 years or less after undergone through surgery. Here if patients survived 5 years or more is represented as 0 and patients who survived less than 5 years is represented as 1.

After collecting all the data the data is splitted into train and test.



**4. METHODOLOGY**

**4.1. EXPLORATORY DATA ANALYSIS:**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns ,to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

4.1.1 FIGURES AND TABLE:

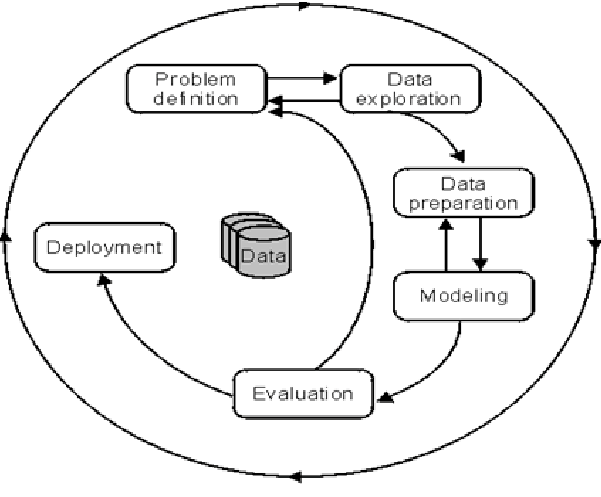


Fig:Data process flow

****

Fig 4.1: Heat Map

**4.1.1. Observation:**

A heat map is a two-dimensional representation of data in which values are represented by colors. A simple heat map provides an immediate visual summary of information. More elaborate heat maps allow the viewer to understand complex data sets.

There can be many ways to display heat maps, but they all share one thing in common – they use color to communicate relationships between data values that would be would be much harder to understand if presented numerically in a spreadsheet.

Fig 4.6: dataset

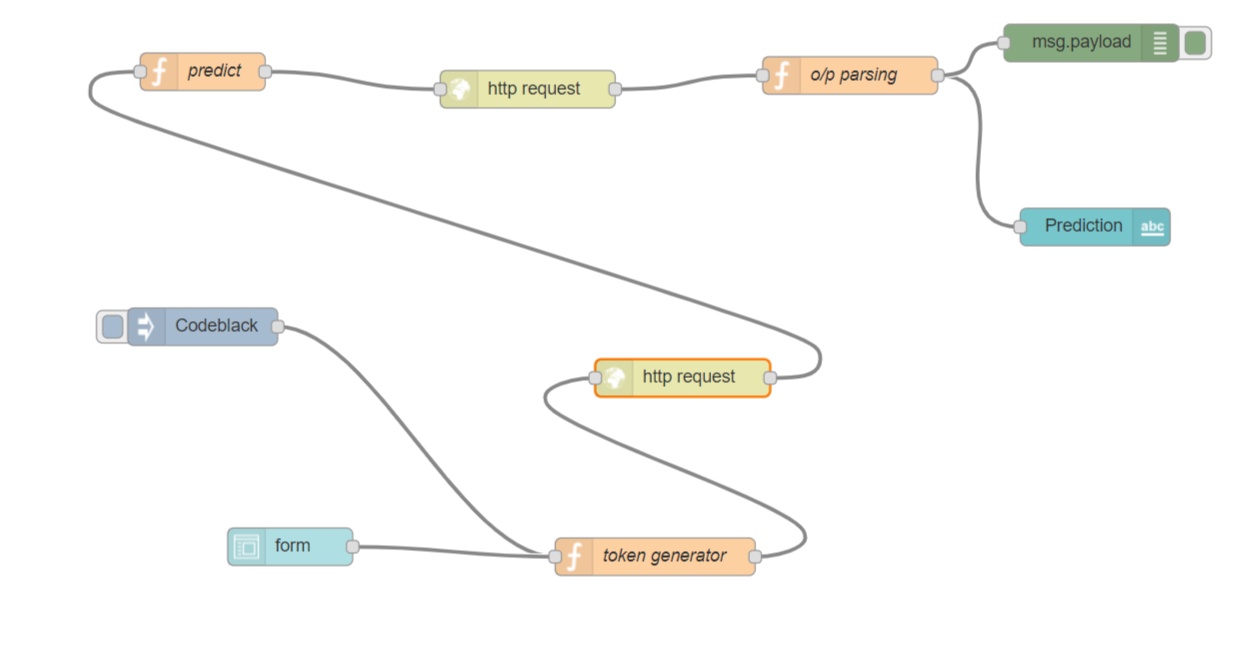
****

Fig:node red

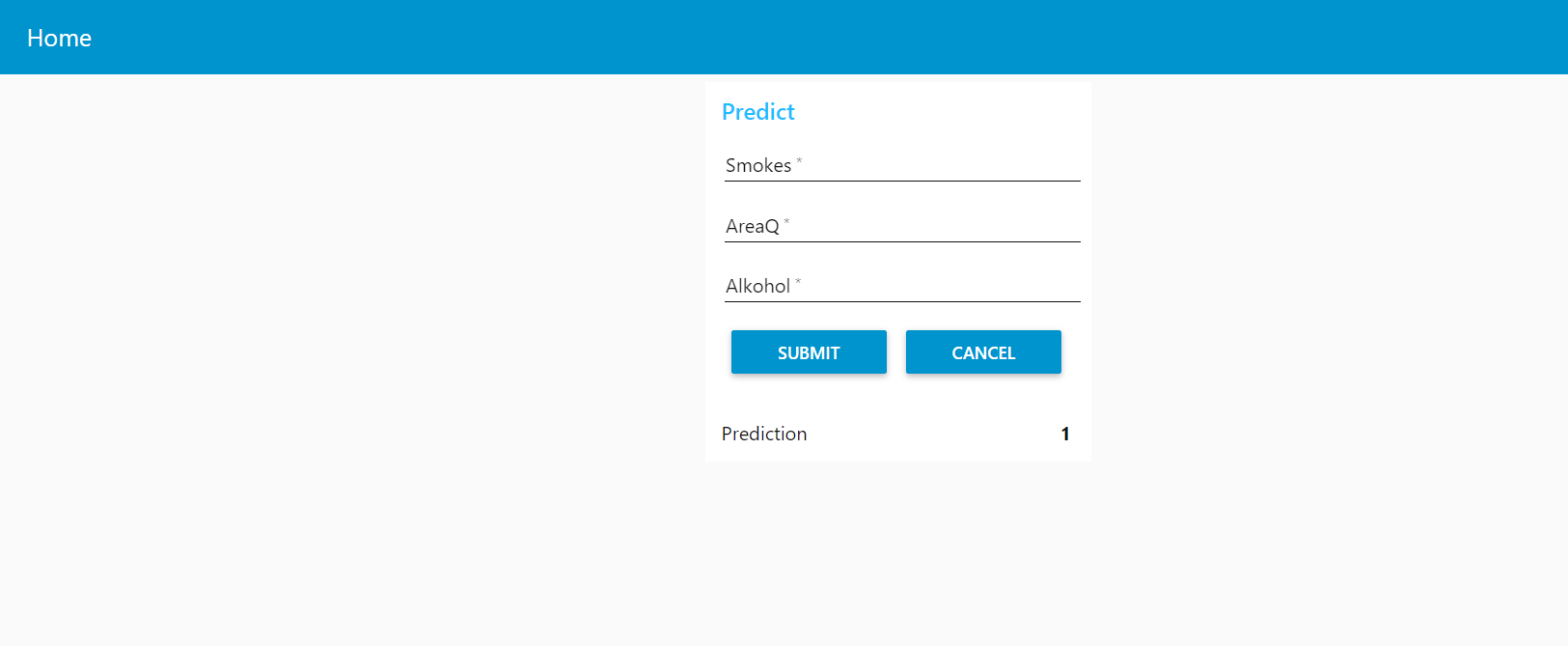


Fig:user interface

**4.2.Data Modelling:**

Since our project is a classification type, we implemented logistic regression. Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values).

**4.2.1. Logistic Function:**

Logistic regression is named for the function used at the core of the method, the logistic function.

The [logistic function](https://en.wikipedia.org/wiki/Logistic_function), also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

Where e is the [base of the natural logarithms](https://en.wikipedia.org/wiki/E_(mathematical_constant)) (Euler’s number or the EXP () function in your spreadsheet) and value is the actual numerical value that you want to transform.

**4.2.2. Representation**

Logistic regression uses an equation as the representation, very much like linear regression.

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modelled is a binary values (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

y = e^ (b0 + b1\*x) / (1 + e^ (b0 + b1\*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b’s).

The logistic regression model takes real-valued inputs and makes a prediction as to the probability of the input belonging to the default class (class 0). If the probability is > 0.5 we can take the output as a prediction for the default class (class 0), otherwise the prediction is for the other class (class 1). For this dataset, the logistic regression has three coefficients just like linear regression, for example:

Output = b0 + b1\*x1+b2\*x2+b3\*x3

The job of the learning algorithm will be to discover the best values for the coefficients (b0, b1 and b2) based on the training data. Unlike linear regression, the output is transformed into a probability using the logistic function:

P (class=0) = 1 / (1 + e^ (-output))

This would be written as:

P (class=0) = 1 / (1 + EXP (-output))

Logistic regression does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the logit of the explanatory variables and the response.

Independent variables can be even the power terms or some other nonlinear transformations of the original independent variables.

The dependent variable does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal…); binary logistic regression assume binomial distribution of the response.

The homogeneity of variance does NOT need to be satisfied.

Errors need to be independent but NOT normally distributed.

It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.

**5.FINDINGS AND SUGGESTIONS**

Apart from logistic regression, decision tree and artificial neural networks also been used for predicting and analyzing

the cancer patients dataset.

Suggestions:

* Modify models to use different metric rather than ROC(auc) which takes in consideration the best threshold
* Try different stacking models

There are also different data mining techniques that can be used for the prediction of lung cancer recurrence.Researchers analyzed lung cancer data using three classification techniques to predict the recurrence of the cancer and then compared the results. The results indicated that SVM are the best classifier predictor with the test dataset, followed by ANN and DT.

**6. CONCLUSION**

This is an analysis of the Haberman’s survival analysis Dataset*,* obtained from Kaggle. We analysed several machine learning classification models to compare their results. The best model found is Logistic regression and reaches with an accuracy score of about 1.0