**PROJECT REPORT**

**1. SURVIVALANALYSIS-CANCER (BREASTCANCER) PREDICTION**

**1.1. Introduction:**

Breast cancer starts when cells in the breast begin to grow out of control. These cells usually form a tumor that can often be seen on an x-ray or felt as a lump. The tumor is malignant (cancer) if the cells can grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. Breast cancer occurs almost entirely in women, but men can get breast cancer, too.

Breast cancer is the most common malignant disease for females and the second most common type of cancer after lung cancer for both sexes. It primarily affects women older than 50 years. Even though the absolute incidence in women aged 20 - 40 years is low, breast cancer constitutes about 24 percent of new cancers in this age group. Hence treatment of breast cancer, including surgery, drugs (hormone therapy and chemotherapy) and radiation, is a main interest of the public health sector .Breast cancer is highly curable if diagnosed at an early stage. Traditional prognostic factors include auxiliary lymph node status, tumour size, nuclear grade and histological grade etc. They are important predictors of whether a tumour is localized and therefore amenable to local treatment. Many researches have been studying the relationship between these clinical variables and the survival time of breast cancer patients.

**1.2 Objective of research:**

This research investigates the survival status of the patient who had undergone surgery for breast 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 breast cancer victim who had undergone surgery.

**1.3 Problem Statement:**

Breast 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 breast 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**

Breast 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 and prognostic factors for breast cancer, there is a need to get an analysis about the breast 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.

**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 breast cancer.

**3.1. Source:**

https://www.kaggle.com/gilsousa/habermans-survival-data-set.

There are 4 attribute in this data set out of which 3 are features and 1 class attribute as below. Also, there are 306 instances of data.

1. Number of Axillary nodes (Lymph Nodes)

2. Age

3. Operation Year

4. Survival Status

**3.2. Lymph Node:**

Lymph nodes are small, bean-shaped organs that act as filters along the lymph fluid channels. As lymph fluid leaves the breast and eventually goes back into the bloodstream, the lymph nodes try to catch and trap cancer cells before they reach other parts of the body. Having cancer cells in the lymph nodes under your arm suggests an increased risk of the cancer spreading. In our data it is axillary nodes detected (0–52)

**3.3. Age:**

It represent the age of patient at which they undergone surgery (age from 30 to 83)

**3.4. Operation year:**

Year in which patient was undergone surgery (1958–1969).

**3.4. 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 1 and patients who survived less than 5 years is represented as 2.

**4. METHODOLOGY**

**4.1. EXPLORATORY DATA ANALYSIS:**

4.1.1 FIGURES AND TABLE:

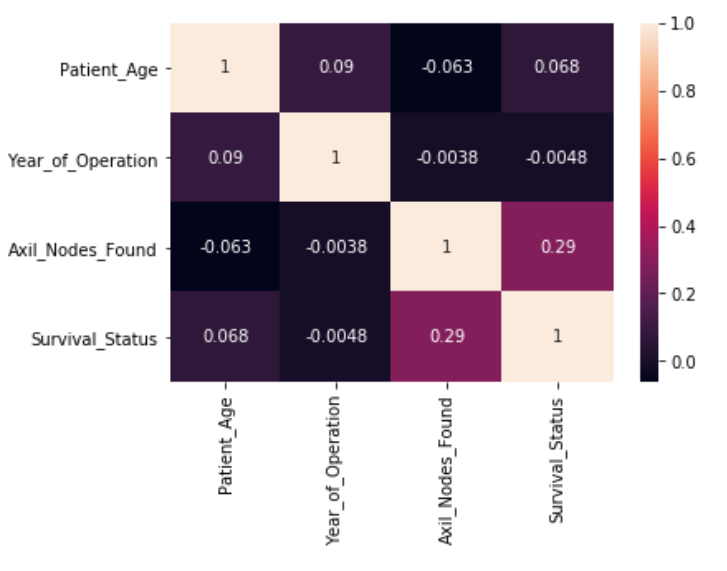


Fig 4.1: Heat Map

**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.

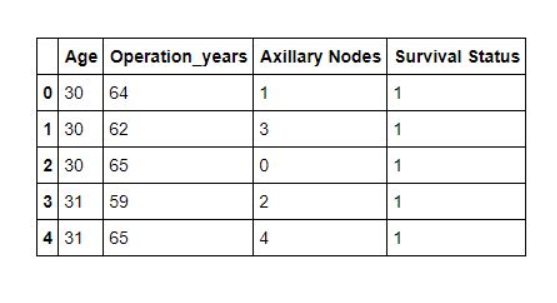
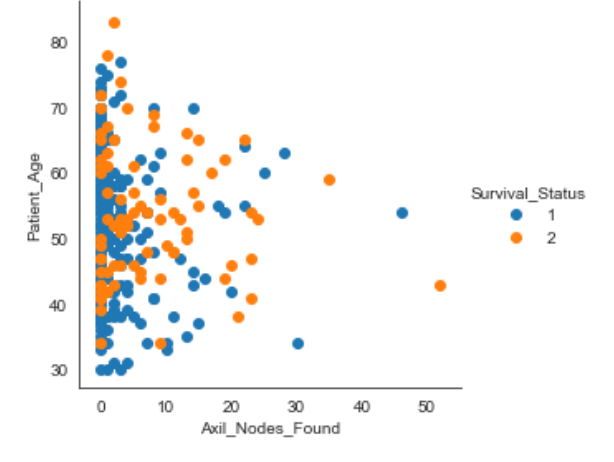


Fig 4.1: Dataset



**Fig 4.2: Graph across Patient\_Age and Axil\_Nodes\_Found**

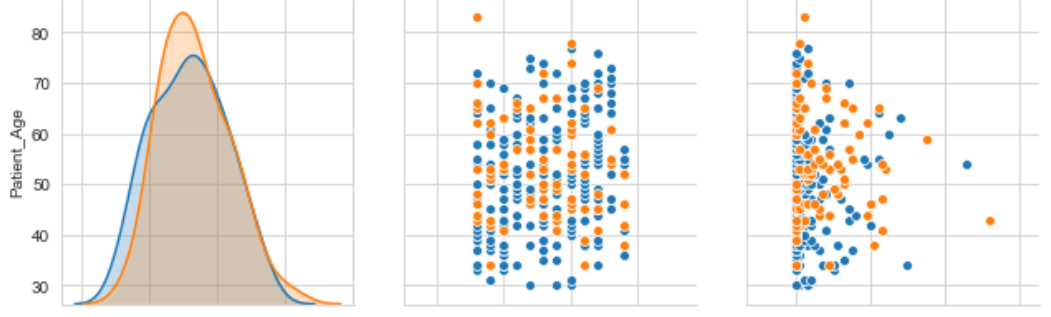


Fig 4.3: Pair plot graphs

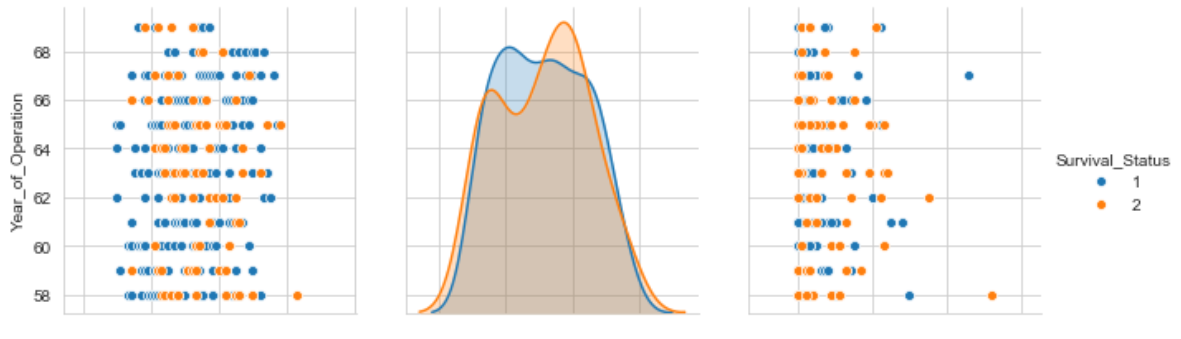


Fig 4.4: Pair plot graphs

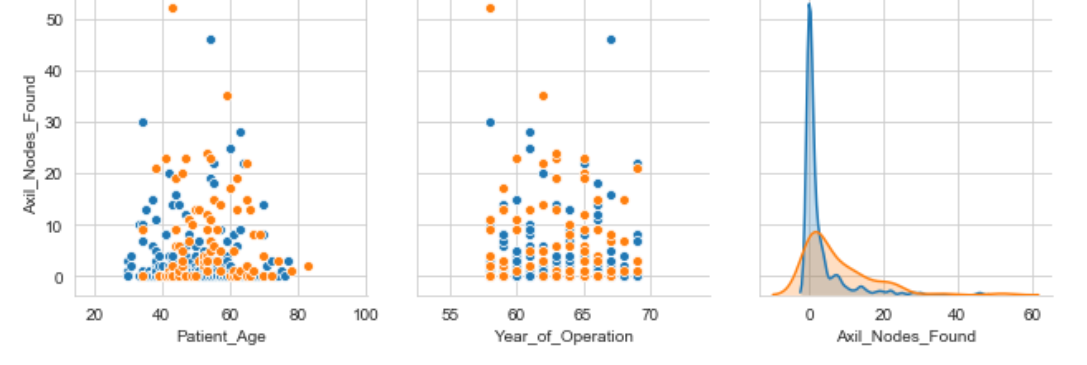


Fig 4.5: Pair plot graphs

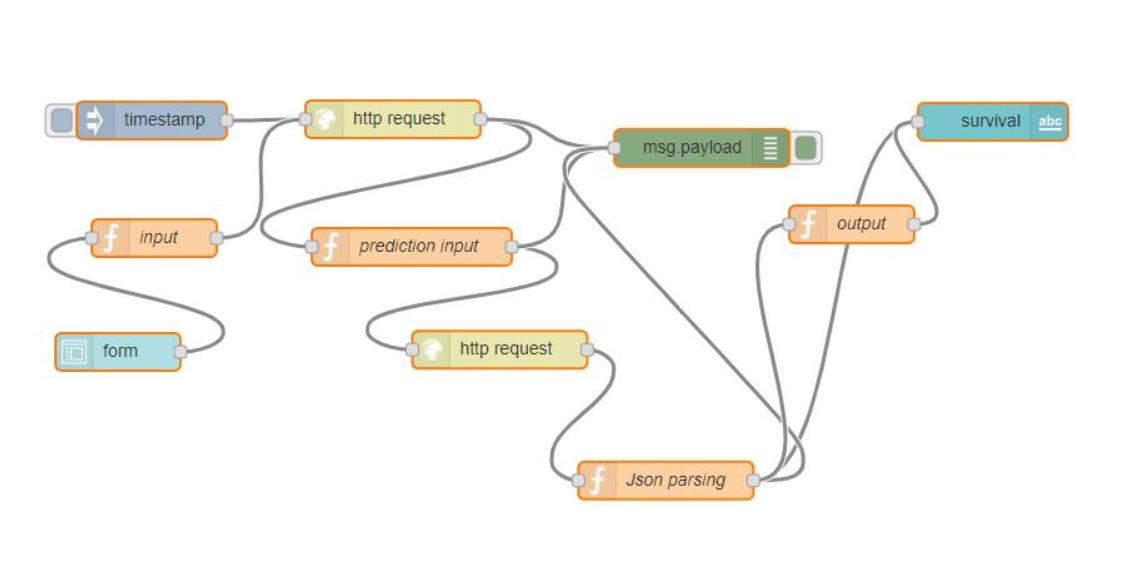


Fig 4.6: Node-red flow



Fig 4.8: User Interface

**4.2Data 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. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.

**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. REFERENCES**

* <http://cs229.stanford.edu/notes/cs229-notes1.pdf>
* <http://machinelearningmastery.com/logistic-regression-for-machine-learning/>
* <https://www.breastcancer.org/symptoms/diagnosis/lymph_nodes>
* [https://github.com/IamRiddhi/Haberman-s-Breast-Cancer-Survival-Prediction](https://www.google.com/url?q=https://github.com/IamRiddhi/Haberman-s-Breast-Cancer-Survival-Prediction&sa=D&source=hangouts&ust=1558757377524000&usg=AFQjCNFx_-ARbBO-QuHX-zK439ltrA1jjA)
* <https://www.kaggle.com/gilsousa/habermans-survival-data-set>

**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 0.739

Next things to try:

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