GraphicEra
(Deemed to be University)
Accredited by NAAC with Grade A

A PROJECT REPORT
ON

BREAST CANCER PREDICTION USING MACHINE LEARNING.

Name	:	Dhruv Saini
University Roll Number	:	2015241
Subject	:	Julia AI Project

Under the guidance of

MR MD FARMANUL HAQUE

ACKNOWLEDGEMENT

We are extremely thankful to our honourable President sir Prof. (Dr.) Kamal Ghanshala of GRAPHIC ERA DEEMED TO BE UNIVERSITY for providing all kind of educational and infrastructural support to work in this project, without which this project would not have been possible.

We would like to thank our head of the department, Dr Devesh Pratap Singh and all other faculties of computer science and engineering department, head of the department and faculties of other departments and all other teaching and nonteaching staff of our collage for the most effective and most valuable guidance.

We hereby like to express our sincere gratitude and respect to our project guide Mr. Md Farmanul Haque for his stimulating guidance and continuous supervision, monitoring and constant encouragement throughout the project completion. The blessing, help and guidance given by him/her time to time shall us a long way in the journey of life on which we are about to embark.

We are obliged to our project team member for the valuable information provided by then in their respective fields. We are grateful for everyone's cooperation during the period of our project assignment.

INDEX

ABSTRACT

Aim: Certification of breast centres helps improve the quality of care but requires additional resources, particularly for documentation. There are currently no published data on the actual staff costs and financial resources required for such documentation. The aim of this study was to determine the time and resources required to document a patient with primary breast cancer from diagnosis to the end of follow-up, to establish a database for future strategic decisions.

Material and Methods: All diagnostic and therapeutic procedures of patients with primary breast cancer were recorded at the University Breast Center of Franconia. All time points for documentation were evaluated using structured interviews. The times required to document a representative number of patients were determined and combined with the staff costs of the different professional groups, to calculate the financial resources required for documentation.

Results: A total of 494 time points for documentation were identified. The study also identified 21 departments and 20 different professional groups involved in the documentation. The majority (54%) of documentation was done by physicians. 62% of all documentation involved outpatients. The results of different scenarios for the diagnosis, therapy and follow-up of breast cancer patients in a certified breast center showed that the time required for documentation can be as much as 105 hours, costing € 4135.

Conclusion: This analysis shows the substantial staffing and financial costs required for documentation in certified centers. A multi-center study will be carried out to compare the costs for certified breast centers of varying sizes with the costs of non-certified care facilities.

1.1. INTRODUCTION

According to information from the Federal Statistical Office of Germany, breast cancer is the most common cause of death from malignant disease in women in Germany. According to figures of the Robert Koch Institute for 2010, 70340 women develop breast cancer every year in Germany. The projection for 2014 is 75200 new cases for that year [1]. Allowing for an annual mortality of 17466 women and taking the 10-year follow-up into account, more than half a million women in Germany are currently receiving care and being followed up. An estimated 60000 women additionally have metastases. To put this into context, approximately 32033000 adult women were living in Germany in 2010 [2], meaning that one in 55 women is receiving oncological care or follow-up for breast cancer. These figures show that documentation and quality assurance for breast cancer patients are extremely relevant for healthcare policies and health economics. Documentation and quality assurance are increasingly central for health economics, both nationally and internationally, and it is important that further studies on these topics are carried out [3]. German healthcare policies have also recognized the importance of obtaining more detailed information and have begun to focus more on this aspect. On June 16, 2008, the German Federal Ministry of Health (BMG) launched the National Cancer Plan [4,5]. The goal of the National Cancer Plan is to optimize the care of cancer patients [6]. Key aspects include improving the provision of oncology services, ensuring guideline-based care and improving quality assurance while ensuring that documentation is both effective and cost-efficient. Cancer is a highly complex disease and it is indisputable that documentation plays an important role during the course of disease and treatment [7]. Moreover, more than for any other disease, cancer treatment involves numerous medical specialties and professional groups from many different care services, and the complexity of the disease means that patients may have to be followed up for the rest of their lives. A reliable and neutral representation of important stations during the course of disease, starting from diagnosis and including treatment and follow-up, is therefore indispensable to optimize care [7]. Tumour documentation is not only necessary when formulating the complex treatment process. In addition to healthcare providers and patients, healthcare insurance companies, researchers and politicians also depend on tumour documentation as a source of reliable information on the quality of oncological care in Germany. However, there are currently a number of problems associated with the quality indicators used to document care for quality assurance purposes. Numerous quality indicators are used to describe the quality of structures, processes and outcomes. But the same indicators are defined differently in different documentation systems. At present, comparisons using different documentation systems are usually not possible (for example, the number of patients with primary breast cancer in a single center: some systems count individual patients, others list each breast separately). There are also many different quality assurance systems in use. Examples for this include the documentation required by the German Cancer Society (DKG) and the German Society for Senology (DGS) from certified breast centers, the data collected by the AQUA Institute, the data required to participate in the Disease Management Program (DMP) Breast Cancer, the data collected for quality assurance purposes in breast cancer screening, etc. The existing documentation systems have very different objectives and describe different aspects of care provision, and data are processed using different systems for data acquisition and reporting.

Certified breast centers provide tertiary care to breast cancer patients and stand in particular need of additional personnel and financial resources to comply with the requirements to document quality parameters.

The tumour documentation of breast cancer patients is thus a prime example of how developments can go wrong. Quality assurance in oncology is currently very heterogeneous in Germany, and clear and uniform guidelines are lacking. Multiple data collections, different documentation systems and partial data collection should be unacceptable; data documentation should always be defined in terms of the goal of the data collection, for example, to certify the quality of services, reinforce compliance with guidelines, improve the process quality and – the most important aspect for patients – improve outcomes. Changes in documentation could reduce the number of staff required and save costs.

But in order to achieve these goals it is necessary to take stock and review the **situation as it currently stands**, including:

- At which points in the patient's course (from the initial diagnosis to treatment, follow-up and quality assurance) are data generally collected?
- Which parameters are documented?
- Who (professional group/medical specialty) does the documenting?
- Which financial resources are currently required for a complete documentation of a breast cancer patient?

To calculate the resources actually needed for documentation, identify which data is collected unnecessarily several times over, and establish where interfaces exist which could be used to optimize data collection across all areas, it is necessary to capture and describe the current status of data collection. The obvious gap between currently available resources for data collection and what would actually be needed has not been previously investigated. This is surprising, considering the costs and time spent on documentation. The aim of this study was to provide a detailed horizontal cross-section of the time, costs and staff involved in documenting a patient with primary breast cancer from initial diagnosis and including treatment, follow-up and quality assurance, and to determine the different time points for data collection and the collected parameters.

1.2. BACKGROUND OF PROJECT

Breast cancer is the most common non-skin cancer among American women. An estimated 271,270 new cases of breast cancer will be diagnosed in women in the United States in 2019, according to the American Cancer Society. Breast cancer accounts for 15 percent of all new cancer diagnoses and 7 percent of all cancer deaths each year.

What causes breast cancer?

Breasts are made of a variety of different tissues, including ducts, lobes and lobules and glands that produce milk and carry it to the nipple. The breasts also contain lymph nodes and fatty tissue. Cancer develops when cells in the breast mutate and grow out of control, forming a tumour. Most breast cancers—about 80 percent—are ductal carcinomas, which begin in milk ducts. About 10 percent of all breast cancers are lobular carcinomas, which develop in the lobes or glands that produce milk.

Other factors that may increase a woman's risk for developing breast cancer include:

- Obesity
- Breast density
- Menstrual history
- A sedentary lifestyle
- Heavy drinking
- Previous medical treatments

Who gets breast cancer?

The risk for developing breast cancer increases with age. According to the National Cancer Institute:

- The average age of a woman diagnosed with breast cancer is 62.
- The average age of a woman who dies from breast cancer is 68.
- Breast cancer is the most common cancer diagnosed in women between age 55 and 64.
- About 10 percent of breast cancers occur in women younger than 45.

Women with a family history of breast cancer may be at a higher risk for developing the disease. For example:

- Women whose mother, sister or daughter has or had breast cancer may have double the risk.
- Women who have inherited mutations in the BRCA1 or BRCA2 gene are at higher risk.

Breast cancer also occurs in men but is very rare. Approximately 2,670 American men will learn they have breast cancer in 2019, the American Cancer Society estimates. Male breast cancer accounts for 1 percent of all breast cancer diagnoses.

Types of breast cancer

There are many types and subtypes of breast cancer, but most are adenocarcinomas of the breast. Adenocarcinoma tumours are found in many common cancers, including prostate, lung and colorectal. These types of tumours form in glands or ducts that secrete fluid. Breast adenocarcinomas form in milk ducts or milk-producing glands called lobules. Each type of breast cancer may be determined based on where in the breast it develops, whether it is considered invasive or non-invasive and whether it is driven by hormones or proteins. Types of breast cancer include:

- Adenoid cystic carcinoma
- Angiosarcoma
- Ductal carcinoma
- Inflammatory breast cancer
- Lobular carcinoma
- Metaplastic carcinoma
- Phyllodes tumour

Subtypes of breast cancer include those driven by specific hormones, such as estrogen, progestogen or the protein HER2. Sixty percent of breast cancers are estrogen-positive. Twenty percent of breast cancers are HER2-positive. Another 20 percent are triple-negative breast cancers, a type of breast cancer that tests negative for estrogen, progesterone and HER2. Triple-negative breast cancer is among the more aggressive forms of the disease.

Breast cancer cells can spread into the lymph nodes in and around the breasts and, from there, travel and form tumours in distant parts of the body. When that occurs, it is called metastatic breast cancer. When it spreads, breast cancer is most often found in the brain, bones, liver and lungs. It is still considered breast cancer even if it is found on other parts of the body.

Breast cancer symptoms

A lump, mass and change in the feel or position of the breast are among the most common symptoms of breast cancer. Other symptoms include:

- Swelling, redness or inflammation
- Changes in the nipple
- Nipple discharge
- Pain in the breast
- Itchy or irritated breasts
- Changes in colour
- Peeling or flaky skin

Diagnosing breast cancer

Tools and tests used to diagnose breast cancer include:

- Lab tests, including advanced genomic testing.
- Biopsy
- Imaging tests, including ultrasound and mammography

Different tests are used to determine whether the breast cancer has metastasized. These tests include:

- Radiofrequency ablation
- Endobronchial ultrasound
- Bone scan

2. **Mechanism:**

2.1 Importing Libraries

- **DataFrames .jl** is a popular Julia library. It provides a set of tools for working with tabular data. Its design and functionality are similar to those of pandas (in Python) and data.frame, data.table and dplyr (in R), making it a great general purpose data science tool. DataFrames.jl is a great general-purpose tool for data manipulation and wrangling.

- **CSV.jl** is another popular Julia library that provides a number of utilities for working with delimited files. It is build to be fast, flexible and to handle file read, write functions.

- **Plots.jl** is a visualization interface and toolset. It sits above other backends, like GR, PyPlot, PGFPlotsX, or Plotly, connecting commands with implementation. If one backend does not support your desired features or make the right trade-offs, you can just switch to another backend with one command. No need to change your code. No need to learn a new syntax

- **GLM .jl** is a Julia library used to create, fit and test machine learning model. To fit a Generalized Linear Model (GLM), use the function,

```
glm(formula, data, family, link)
```

where,

a) formula: uses column symbols for the DataFrame data.
b) data: a DataFrame which may contain NA values, any rows with NA values are ignored.
c) family: chosen from Bernoulli(), Binomial(), Gamma(), Normal(), or Poisson()
d) link: chosen from the list below, for example, LogitLink() is a valid link for the Binomial() family

2.2 Dataset Analysis

Predicting if the cancer diagnosis is benign or malignant based on several observations/features

1. 30 features are used, examples:
   - radius (mean of distances from centre to points on the perimeter)
   - texture (standard deviation of gray-scale values)
   - perimeter
   - area
   - smoothness (local variation in radius lengths)
   - compactness (perimeter^2 / area - 1.0)
   - concavity (severity of concave portions of the contour)
   - concave points (number of concave portions of the contour)
   - symmetry
2. fractal dimension ("coastline approximation" - 1)
3. Datasets are linearly separable using all 30 input features
4. Number of Instances: 569
5. Class Distribution: 212 Malignant, 357 Benign
6. Target class:
7. Malignant and Benign

2.3 EDA

1. EDA(Exploratory Data Analysis) is essentially a type of storytelling for statisticians . It allows to uncover patterns and insights , often with visual methods , within data . EDA is often the first step of the data modelling process. It is used for gaining a better understanding of data aspects like: - main features of data, -variables and relationships that holds between them, - identifying which variables are important . Various exploratory data analysis methods like:
2. Descriptive Statistics which is a way of a giving a brief overview of the dataset which is dealing with, including some measure and features of the sample DataFrames provide describe() applies basic statistical computations on the dataset like extreme values , count of data , point standard deviation etc. Describe() function gives a group picture of distribution of data.
3. Grouping data an interesting measure which figure out effect of different categorical attributes on other data variables.
4. Anova stands for Analysis of Variance. It is performed to figure out the relation between the different group of categorical data.
5. Under Anova have two measures as result:
6. Correlation and Correlation computation is a simple relationship between two variables in a context such that one variable affects the other. Correlation is different from act of causing. One way to calculate correlation among variables is to find Pearson correlation Here two parameters namely, Pearson coefficient and p-value . There is a strong correlation between two variables when Pearson coefficient is close to either 1 or -1 and the p-value is less than 0.0001.

2.4 Feature Engineering

1. Feature Engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work . Feature engineering is fundamental to the machine learning and is both difficult and expensive . The feature engineering process is
    a. Brainstorming or testing feature ;
    b. Deciding what feature to create;
    c. Creating features;
    d. Checking how the features work with the model;
    e. Improving features if needed;
    f. Go back to brainstorming / creating more features until the work is done .
    g. Preparing the proper input dataset , compatible with machine learning algorithm requirements;
    h. Improving the performance of machine learning .
2. This metric is very impressive to show the importance of feature engineering in data science .
3. Techniques are listed as: IMPUTATION , HANDLING OUTLIERS , BINNING, LOG TRANSFORM, ONE HOT ENCODER, GROUPING OPERATIONS, FEATURE SPLIT, SCALING, EXTRACTING DATE.
4. In machine learning and pattern recognition , a feature is an individual measurable property or characteristic of a phenomenon being observed . Choosing informative , discriminating and independent features is a crucial step for effective algorithms in pattern recognition , classification and regression. While feature engineering requires label times, in our general-purpose framework, it is *not hard-coded* for specific labels corresponding to only one prediction problem. If we wrote our feature engineering code for a single problem — as feature engineering is traditionally approached — then we would have to redo this laborious step every time the parameters change.
5. Instead, we use APIs like Featuretools that can build features for *any set of labels without requiring changes to the code.* This means for the customer churn dataset, we can solve multiple prediction problems — predicting churn every month, every other week, or with a lead time of two rather than one month — using the exact same feature engineering code.

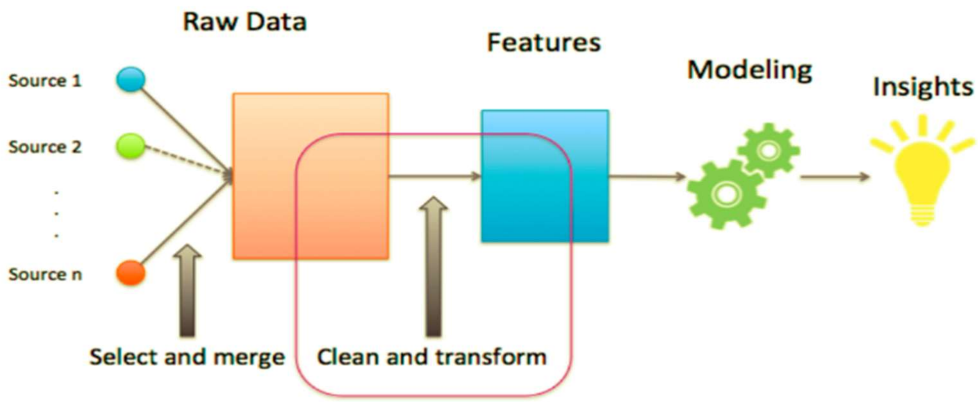


Fig: the workflow of feature engineering

6. Feature engineering, the second step in them machine learning pipelines, takes in the label times from the first step — prediction engineering — and a raw dataset that needs to be refined. Feature engineering means building features for each label while *filtering the data used for the feature based on the label's cutoff time* to make valid features. These features and labels are then passed to modeling where they will be used for training a machine learning algorithm.

2.5 Feature Scaling

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range . It is present in the data in a fixed range . It is performed during the data pre-processing to handle highly varying magnitudes or values or units . If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values , regardless of the unit of the values.

Techniques are:

1. MIN-MAX NORMALIZATION: This techniques re-scales a feature or observation Value with distribution value between 0 and 1.
2. STANDARDIZATION: It is a very effective technique which re-scales a feature value so that it has distribution with 0 and mean value and variance equals to 1 . Prediction of the class of new data point : The model calculates the distance of this datapoint from centroid of each class group. Finally this data point will belong to that class , which will have a minimum centroid distance from it .

The distance can be calculated between centroid and data point using these methods:

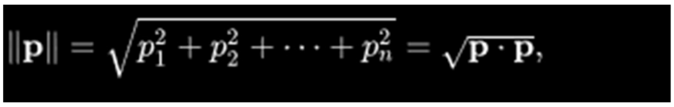
a. EUCLIDEAN DISTANCE: It is the square -root of the sum of squares of differences between the coordinates of datapoint and centroid of each class . The **Euclidean distance** between points **p** and **q** is the length of the line segment connecting them

length of the line segment connecting them

In Cartesian coordinates, if **p** = (p_1, p_2,..., p_n) and **q** = (q_1, q_2,..., q_n) are two points in Euclidean n-space, then the distance (d) from **p** to **q**, or from **q**to **p** is given by the Pythagorean formula

$$\begin{aligned} d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) &= \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2} \\ &= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}. \end{aligned} \qquad (1)$$

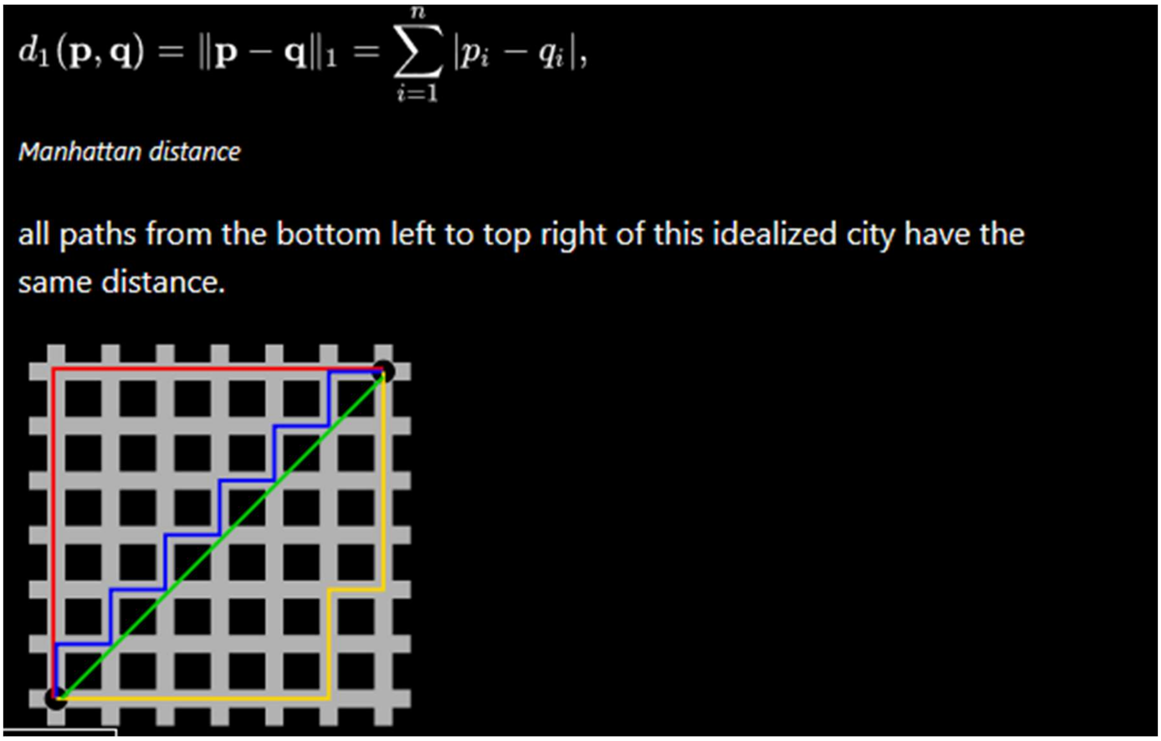
The position of a point in a Euclidean *n*-space is a Euclidean vector. So, **p** and **q** may be represented as Euclidean vectors, starting from the origin of the space (initial point) with their tips (terminal points) ending at the two points. The Euclidean norm, or **Euclidean length**, or **magnitude** of a vector measures the length of the vector:



$$\|\mathbf{p}\| = \sqrt{p_1^2 + p_2^2 + \cdots + p_n^2} = \sqrt{\mathbf{p}\cdot\mathbf{p}},$$

where the last expression involves the dot product.

b. MANHATTAN DISTANCE: It is calculated as the sum of absolute differences between the coordinates of data point and centroid of each class. The distance between two points measured along axes at right angles. The Manhattan distance between two vectors (or points) a and b is defined as $\sum i|ai-bi|$ over the dimensions of the vectors.



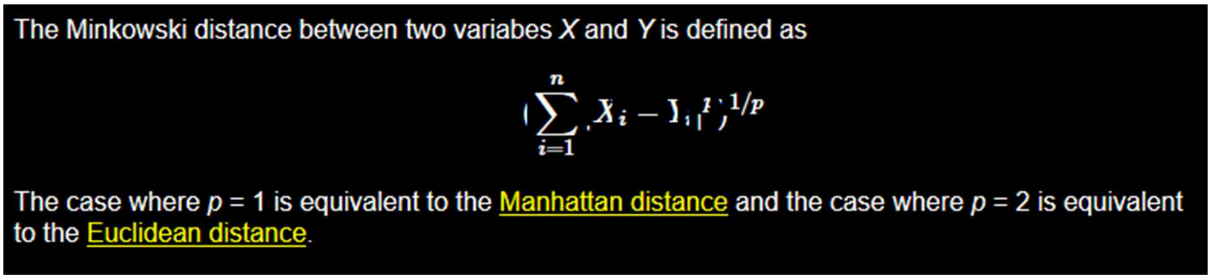
$$d_1(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$

Manhattan distance

all paths from the bottom left to top right of this idealized city have the same distance.

Fig: showing the Manhattan distance between the two points.

c. MINKOWSKI DISTANCE: It is a generalization of two above methods.



The Minkowski distance between two variabes *X* and *Y* is defined as

$$\left(\sum_{i=1}^{n}|X_i - Y_i|^p\right)^{1/p}$$

The case where *p* = 1 is equivalent to the Manhattan distance and the case where *p* = 2 is equivalent to the Euclidean distance.

Although *p* can be any real value, it is typically set to a value between 1 and 2. For values of *p* less than 1, the formula above does not define a valid distance metric since the triange inequality is not satisfied.

2.6 Logistic Regression

Logistic Regression is a classification algorithm used to assign observations to a discrete set of classes . Logistic Regression transforms its output using the logistic sigmoid function to return a probability value . There are two types are: BINARY AND MUILTI – LINEAR FUNCTION FAILSCLASS LOGISTIC REGRESSION . It is used for the classification problems , it is a predictive analysis algorithm and based on the concept of probability .The hypothesis of logistic regression tends it to limit the cost function between o and 1 . Therefore linear function fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression . We can call a Logistic Regression a linear Regression model but the logistic regression uses a more complex cost function can be defined as the '*SIGMOID FUNCTION'* or also known as the 'logistic function' instead of a linear function .It is used to map predictions to probabilities . The func. maps any real value into another value between 0 and 1 .t is a technique to analyse a data-set which has a dependent variable and one or more independent variables to predict the outcome in a binary variable, meaning it will have only two outcomes.

The dependent variable is **categorical** in nature. Dependent variable is also referred as **target variable** and the independent variables are called the **predictors**.

Logistic regression is a special case of linear regression where we only predict the outcome in a categorical variable. It predicts the probability of the event using the log function.

We use the **Sigmoid function/curve** to predict the categorical value. The threshold value decides the outcome(win/lose).

Linear regression equation: $\mathbf{y = \beta0 + \beta1X1 + \beta2X2 \ldots. + \beta nXn}$

- Y stands for the dependent variable that needs to be predicted.
- β0 is the Y-intercept, which is basically the point on the line which touches the y-axis.
- β1 is the slope of the line (the slope can be negative or positive depending on the relationship between the dependent variable and the independent variable.)
- X here represents the independent variable that is used to predict our resultant dependent value.

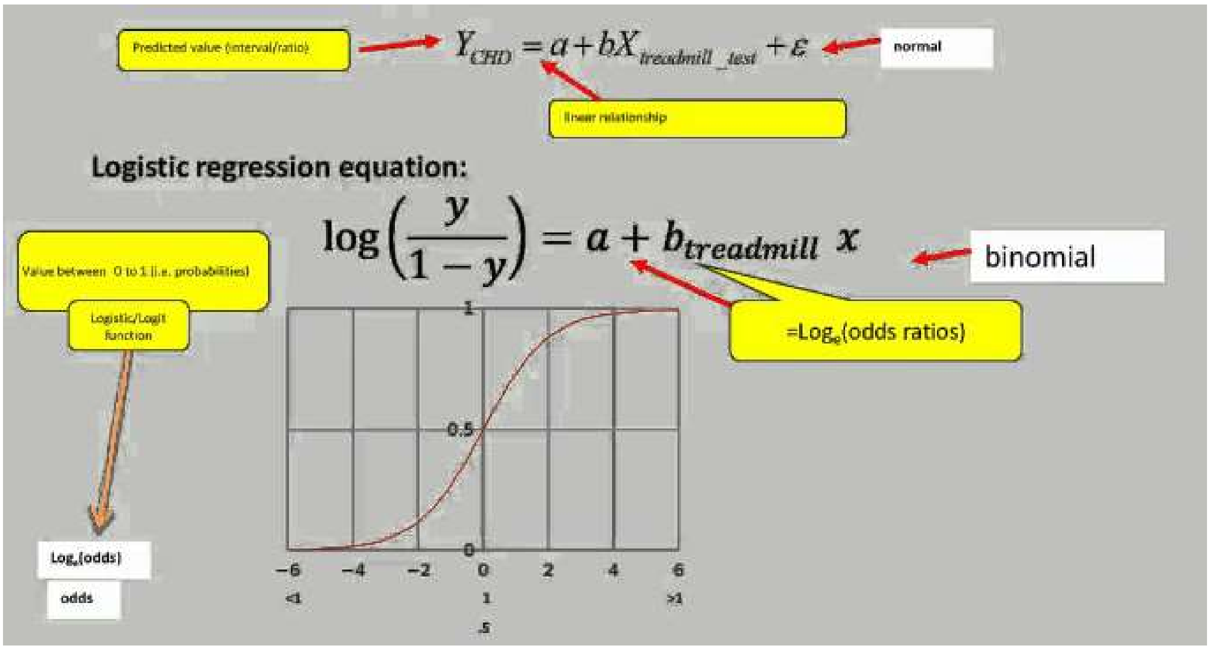
Sigmoid function: $\boldsymbol{p} \mathbf{= 1 / 1 + e^{-y}}$

Logistic Regression equation: $\boldsymbol{p} \mathbf{= 1 / 1 + e^{-(\beta0 + \beta1X1 + \beta2X2 \ldots. + \beta nXn)}}$

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a **logistic** function. It uses a black box function to understand the relation between the categorical dependent variable and the independent variables. Assumptions of Linear Regression. Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions: **Linear relationship. Multivariate normality. No or little multicollinearity**. No auto-correlation. Homoscedasticity.

Logistic regression is **considered** a generalized **linear model** because the outcome always depends on the sum of the inputs and parameters. Or in other words, the output cannot depend on the product (or quotient, etc.) ... "A statistician calls a **model** "**linear**" if the mean of the response is a **linear** function of the parameter, and this is clearly violated for **logistic regression.**

.



Predicted value (interval/ratio)

$Y_{CHD} = a + bX_{treadmill\_test} + \varepsilon$

normal

linear relationship

Logistic regression equation:

$\log\left(\frac{y}{1-y}\right) = a + b_{treadmill}\ x$

binomial

Value between 0 to 1 (i.e. probabilities)

Logistic/Logit function

=Log$_e$(odds ratios)

Log$_e$(odds)

odds

Consider a model with two predictors, and , and one binary (Bernoulli) response variable , which we denote . We assume a linear relationship between the predictor variables, and the log-odds of the event that .

This linear relationship can be written in the following mathematical form (where ℓ is the log-odds, is the base of the logarithm, and are parameters of the model):

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

We can recover the odds by exponentiating the log-odds:

$$\frac{p}{1-p} = b^{\beta_0+\beta_1 x_1+\beta_2 x_2}.$$

.

By simple algebraic manipulation, the probability that is

$$p = \frac{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{b^{\beta_0+\beta_1 x_1+\beta_2 x_2} + 1} = \frac{1}{1 + b^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}.$$

.

The above formula shows that once are fixed, we can easily compute either the log-odds that for a given observation, or the probability that for a given observation. The main use-case of a logistic model is to be given an observation , and estimate the probability that . In most applications, the base of the logarithm is usually taken to be e. However in some cases it can be easier to communicate results by working in base 2, or base 10.

We consider an example with , and coefficients , , and . To be concrete, the model is

$$\log_{10} \frac{p}{1-p} = \ell = -3 + x_1 + 2x_2$$

where is the probability of the event that .Y =1 .

This can be interpreted as follows:

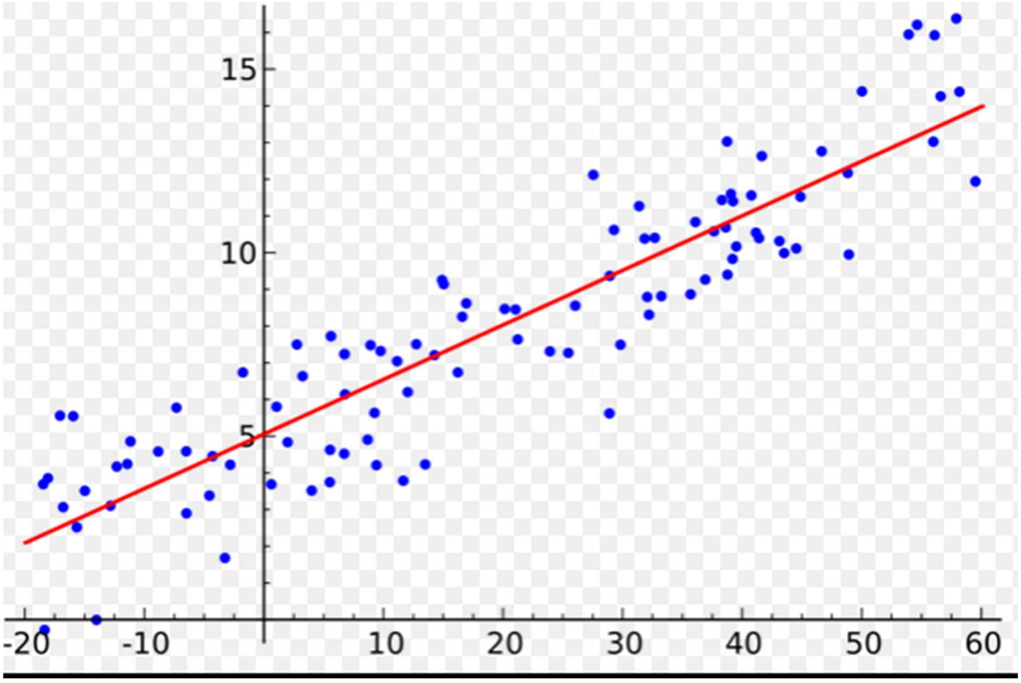
- is the *y*-intercept. It is the log-odds of the event that , when the predictors . By exponentiating, we can see that when the odds of the event that are 1-to-1000, or . Similarly, the probability of the event that when can be computed as .
- means that increasing by 1 increases the log-odds by . So if increases by 1, the odds that increase by a factor of .
- means that increasing by 1 increases the log-odds by . So if increases by 1, the odds that increase by a factor of Note how the effect of on the log-odds is twice as great as the effect of , but the effect on the odds is 10 times greater.

Logistic regression can be seen as a special case of **generalized linear model** and thus analogous to **linear regression**. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of **linear regression**.

Multivariate logistic regression is like simple **logistic regression** but with multiple predictors. **Logistic regression** is similar to linear **regression** but you can use it when your response variable is binary. As in linear regression let's represent our hypothesis(Prediction Of Dependent Variable) in classification. In classification our hypothesis representation which tries to predict the binary outcome of either o or 1, will look like,

$\boldsymbol{h\theta(x) = g(\theta\ T x) = 1/ 1 + e -\theta\ T x}$ **,**

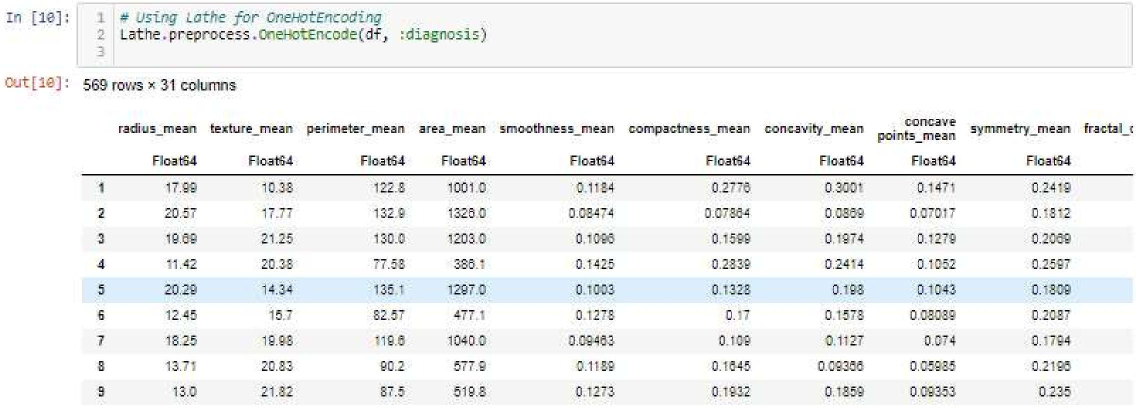
Here g(z) = 1/(1 + e ^−z), is called the l***ogistic function or the sigmoid function.***



here dots are scatterplot suggest the form and strength of the relationship between the dependent variable and regressors.

2.7 One Hot Encoder

A one hot encoding is a **representation of categorical variables as binary vectors**. Each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.



```
In [10]: # Using Lathe for OneHotEncoding
         Lathe.preprocess.OneHotEncode(df, :diagnosis)

Out[10]: 569 rows × 31 columns
```

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_(
	Float64	Float64	Float64	Float64	Float64	Float64	Float64	Float64	Float64	
1	17.99	10.38	122.8	1001.0	0.1184	0.2776	0.3001	0.1471	0.2419	
2	20.57	17.77	132.9	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	
3	19.69	21.25	130.0	1203.0	0.1096	0.1599	0.1974	0.1279	0.2069	
4	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	
5	20.29	14.34	135.1	1297.0	0.1003	0.1328	0.198	0.1043	0.1809	
6	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	
7	18.25	19.98	119.6	1040.0	0.09463	0.109	0.1127	0.074	0.1794	
8	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	
9	13.0	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	

Onehotencoding is a powerful technique to transform categorical data into a numerical representation that machine learning algorithms can utilize to perform optimally without falling into the misrepresentation issue previously mentioned. You should now be able to easily perform **one-hotencoding using** the **Lathe** built-in functionality. Using categorical data in Multiple Regression Models is a powerful method to include non-numeric data types into a regression model. Categorical data refers to data values which represent categories - data values with a fixed and unordered number of values, for instance gender (male/female) or season (summer/winder/spring/fall). In a regression model, these values can be represented by dummy variables- variables containing values such as 1 or 0 representing the presence or absence of the categorical value. The Dummy Variable trap is a scenario in which the independent variables are multicollinear- a scenario in which two or more variables are highly correlated; in simple terms one variable can be predicted from the others. The ***DUMMY VARIABLE TRAP*** is a scenario in which the independent variables are multicollinear – a scenario in which two or more variables are highly correlated .In statistics, especially in regression models, we deal with various kind of data. The data may be quantitative (numerical) or qualitative (categorical). The numerical data can be easily handled in regression models but we can't use categorical data directly, it needs to be transformed in some way.

For transforming categorical attribute to numerical attribute, we can use label encoding procedure (label encoding assigns a unique integer to each category of data). But this procedure is not alone that much suitable, hence, ***One hot encoding*** is used in regression models following label encoding. This enables us to create new attributes according to the number of classes present in the categorical attribute i.e. if there are *n* number of categories in categorical attribute, *n* new attributes will be created. These attributes created are called ***Dummy Variables***. Hence, dummy variables are "proxy" variables for categorical data in regression models.

These dummy variables will be created with *one hot encoding* and each attribute will have value either 0 or 1, representing presence or absence of that attribute.

3.0 Data Visualization

3.1 **Plot.jl** for Julia Data Visualization

It is a high-level data visualization library of Julia that enable a user to create various statistical graphs. If one backend does not support your desired features or make the right trade-offs, you can just switch to another backend with one command.

Plot Type – Scatter Plot

Scatter plots are the most basic plots in exploratory analytics. They help the analyst get a rough idea of the data distribution and the relationship between the corresponding columns, which in turn helps identify some prominent patterns in the data.

Format: `scatter(x, y, title = "My Scatter Plot")`

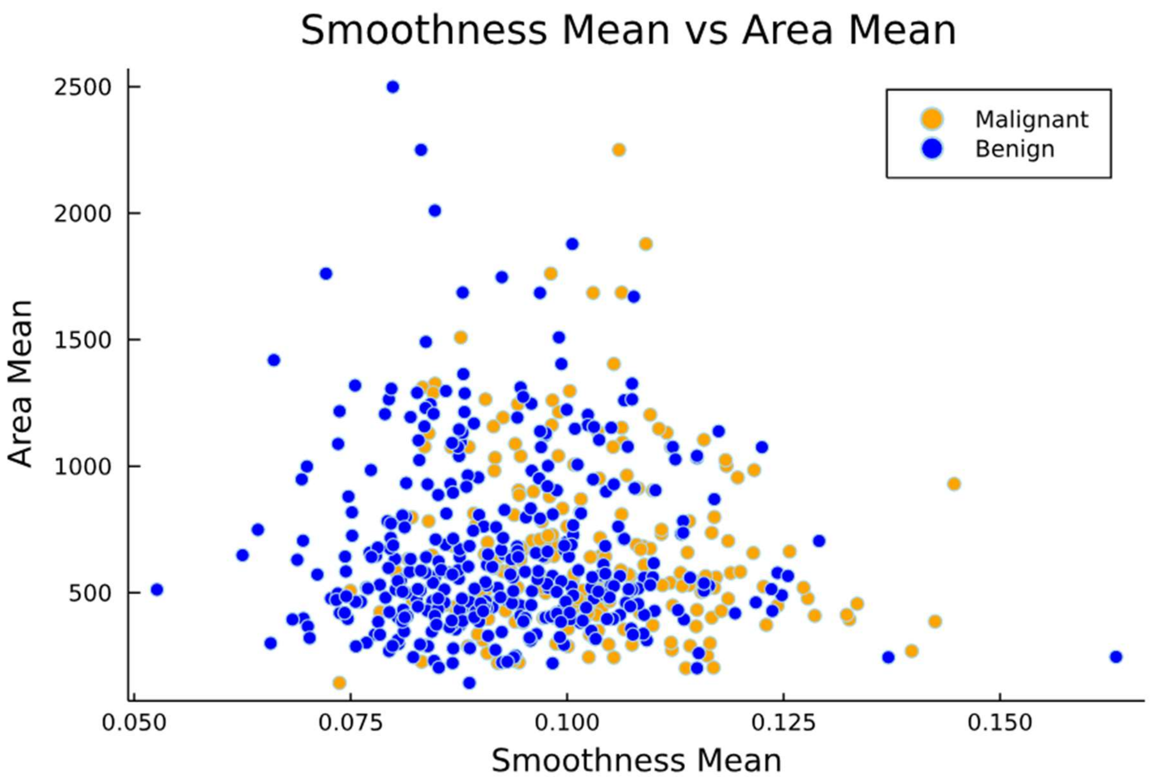


Fig: Smoothness Mean vs Area Mean Plot

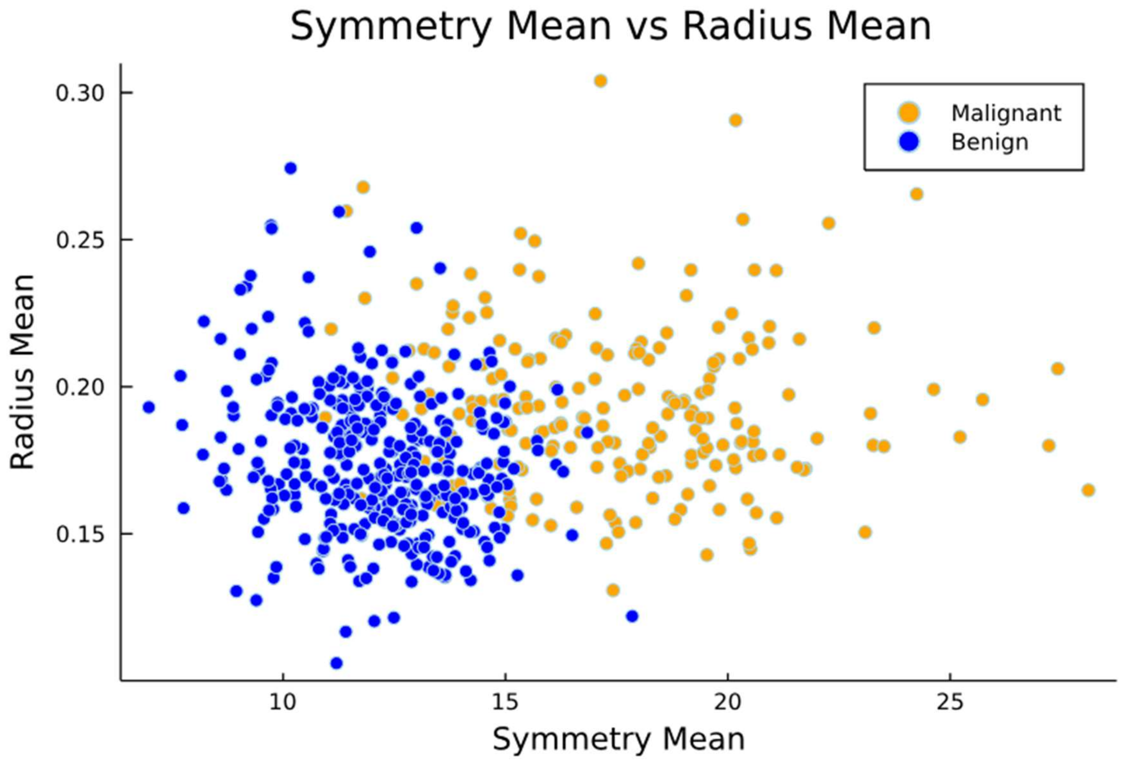


Fig: Symmetry Mean vs Radius Mean Plot

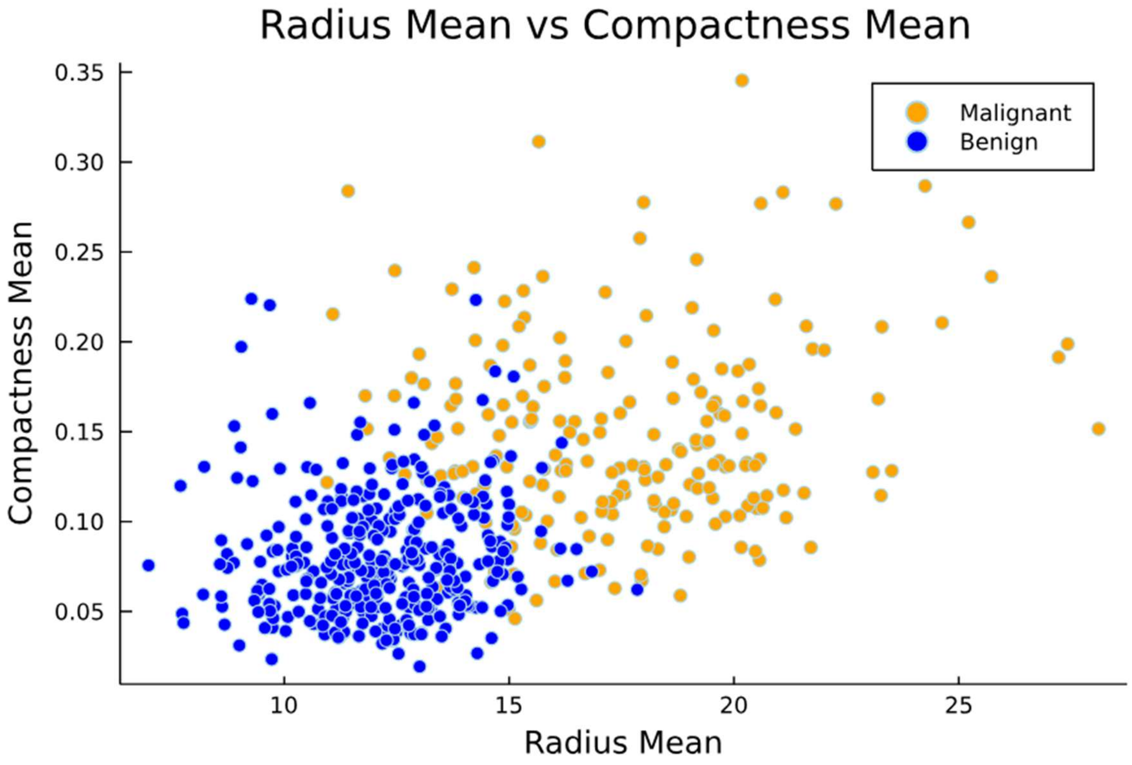


Fig: Radius Mean vs Compactness Mean Plot

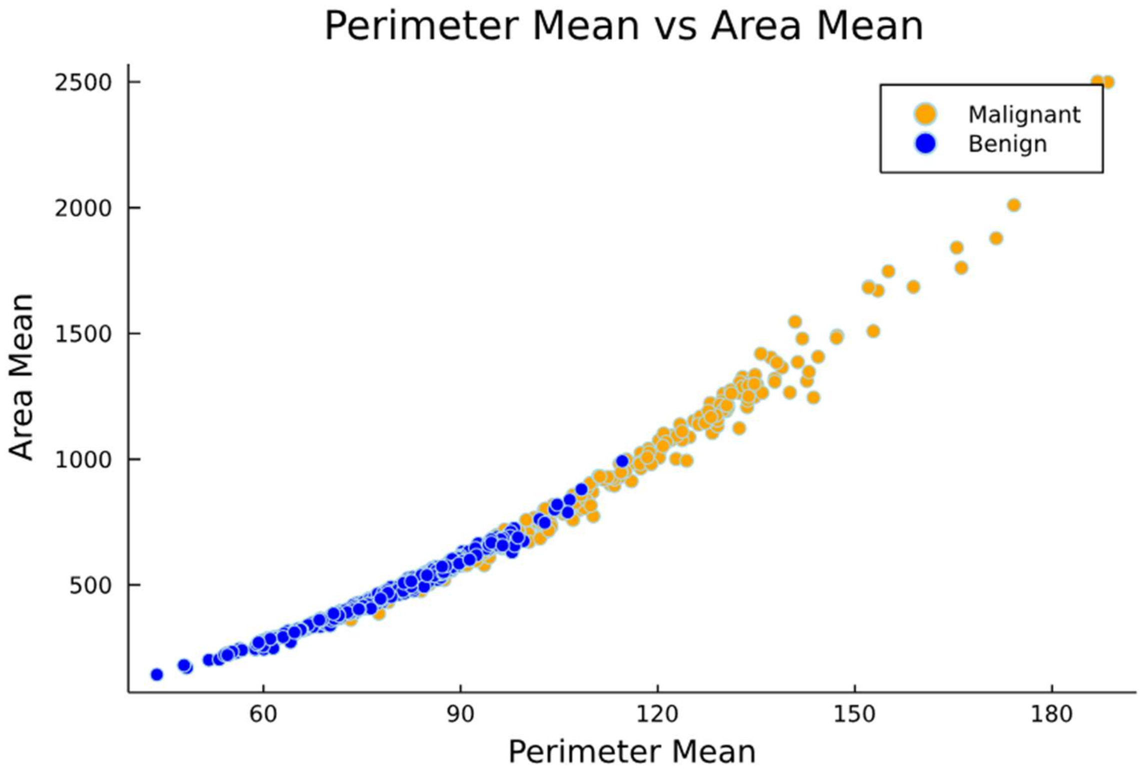


Fig: Perimeter Mean vs Area Mean Plot

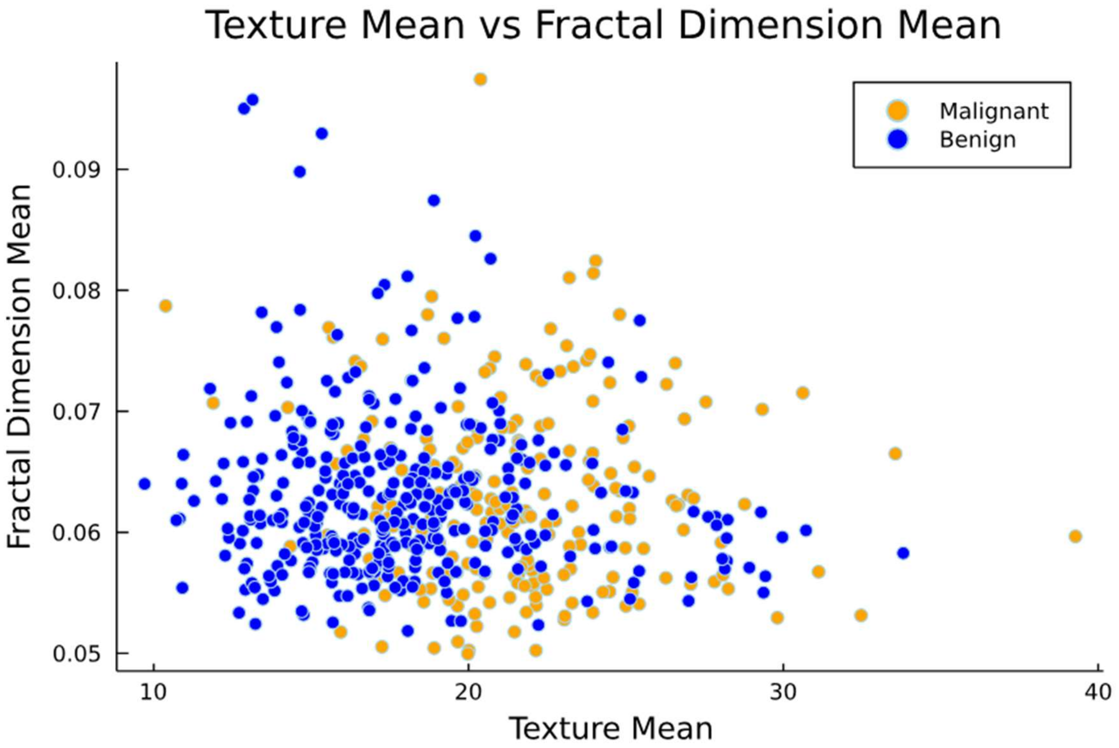


Fig: Texture Mean Plot vs Fractal Dimension Mean

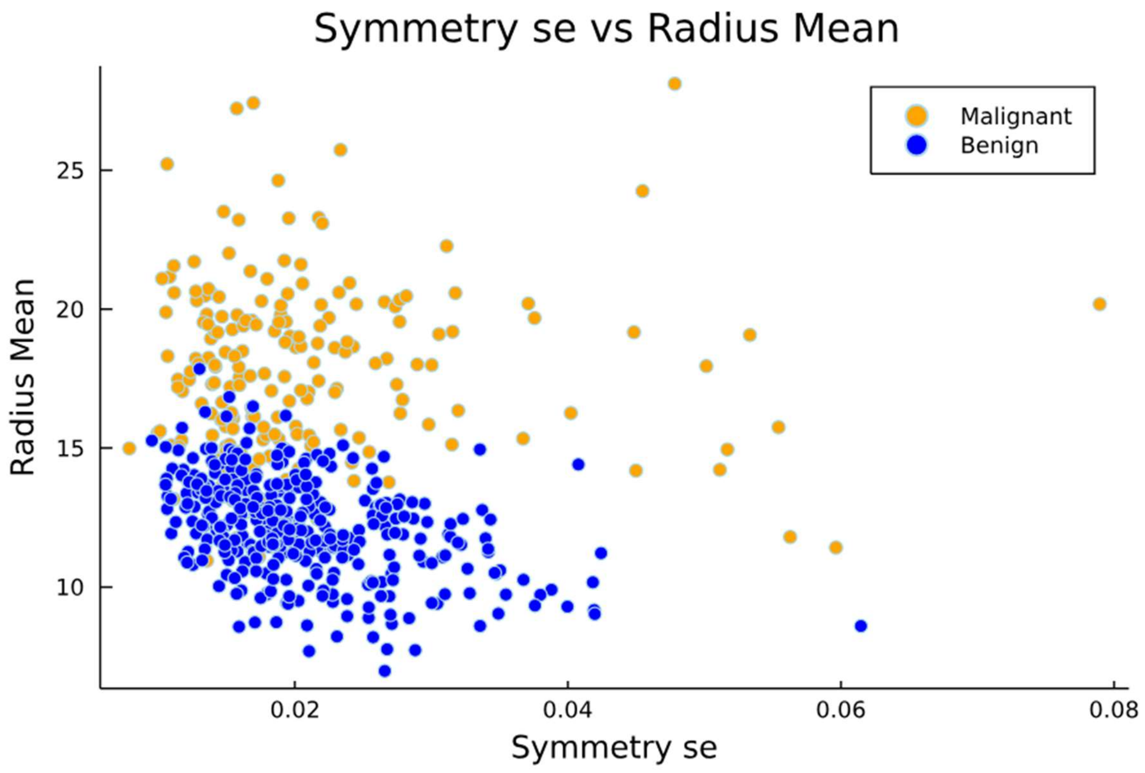


Fig: Symmetry se vs Radius Mean Plot

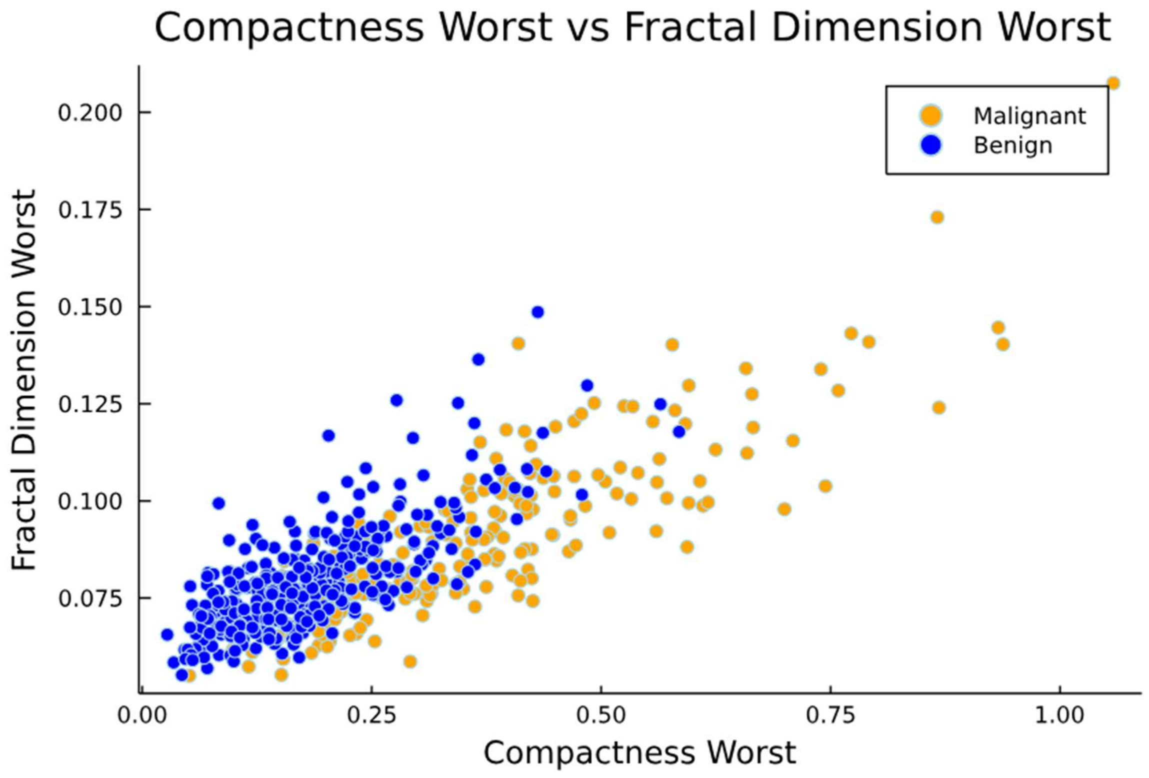


Fig: Compactness Worst vs Fractal Dimension Worst Plot

Plot Type – Histogram Plot

A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.

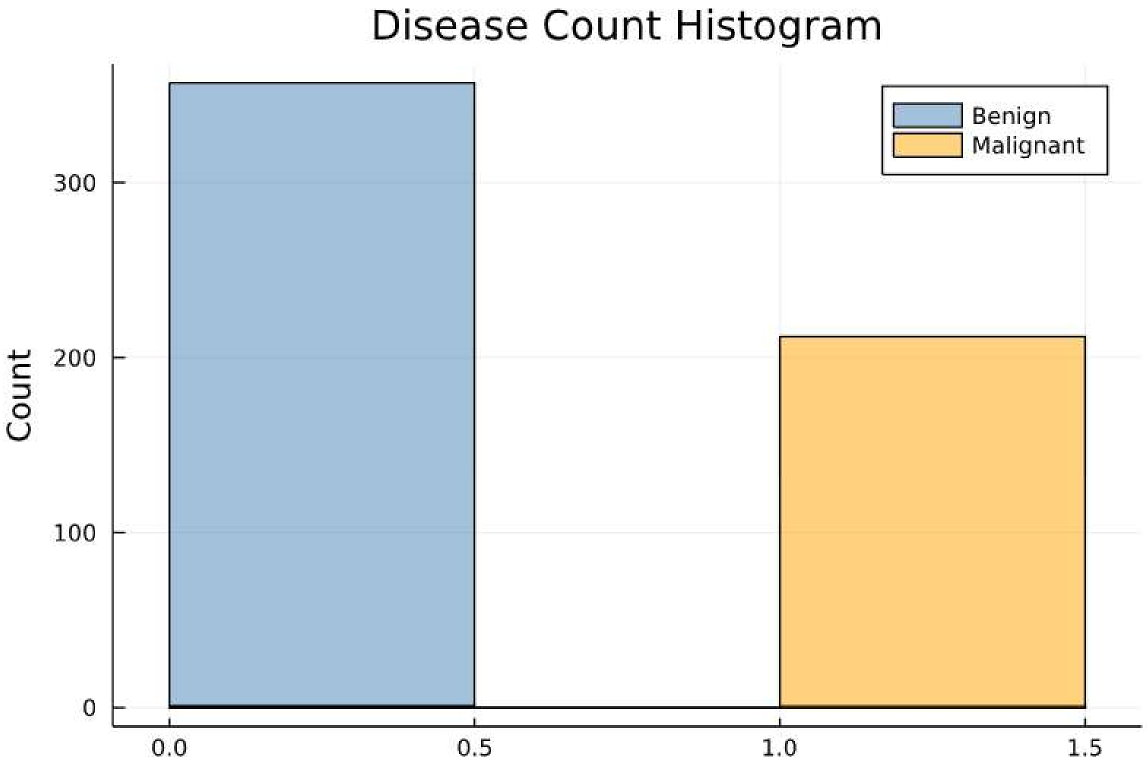


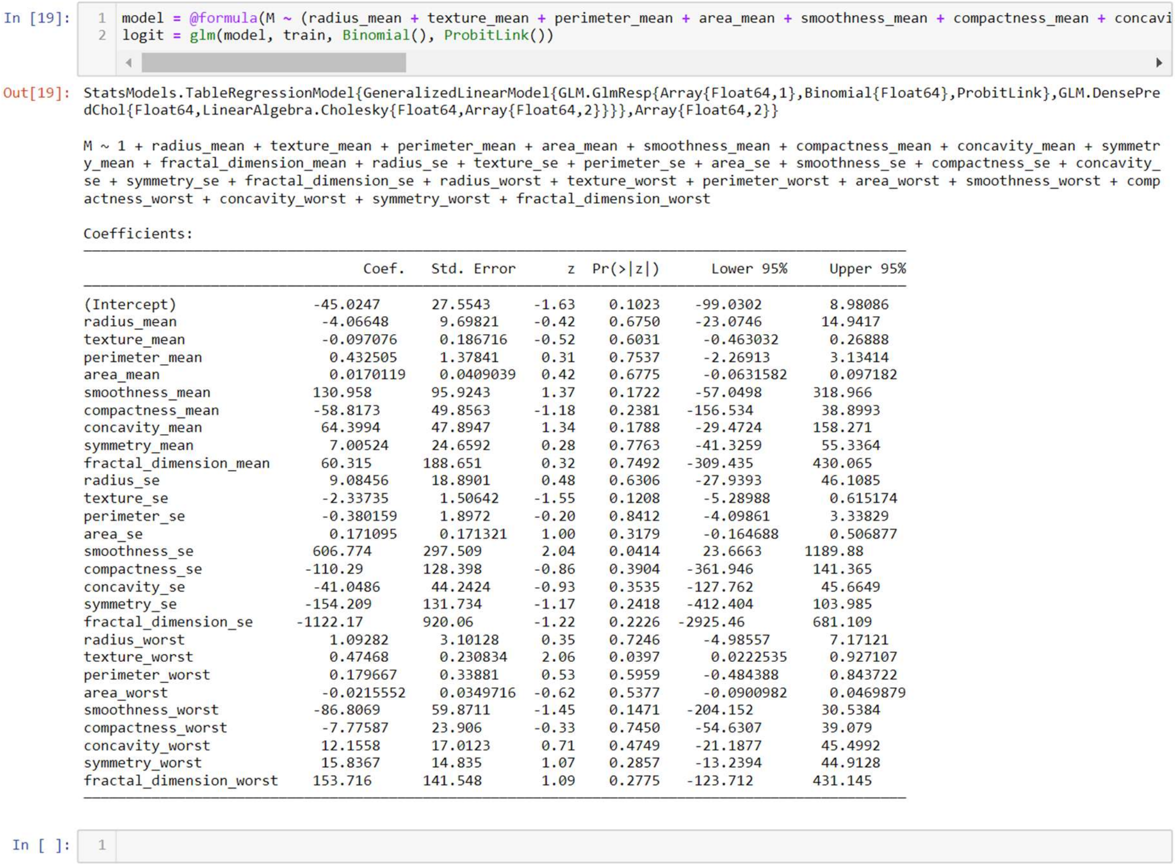
Fig: Disease Count Histogram

4.0 Model

4.1 Training Model

The process of training an ML model involves providing an ML algorithm (that is, the *learning algorithm*) with training data to learn from. The term *ML model* refers to the model artifact that is created by the training process.

The training data must contain the correct answer, which is known as a *target* or *target attribute*. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.



In [19]:

```
model = @formula(M ~ (radius_mean + texture_mean + perimeter_mean + area_mean + smoothness_mean + compactness_mean + concavi
logit = glm(model, train, Binomial(), ProbitLink())
```

Out[19]:

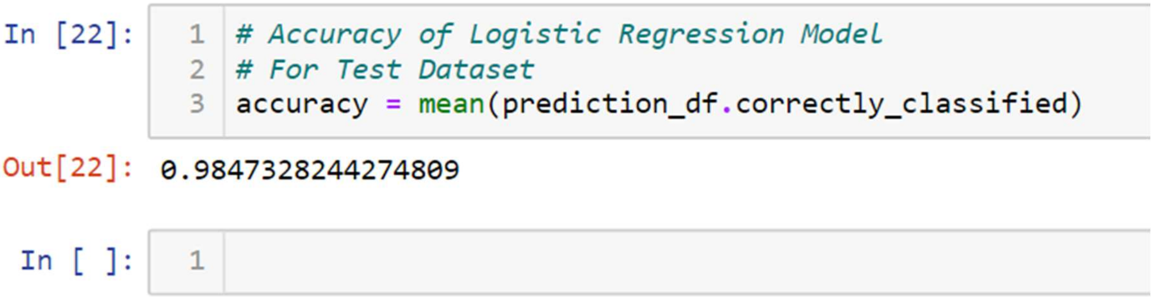
```
StatsModels.TableRegressionModel{GeneralizedLinearModel{GLM.GlmResp{Array{Float64,1},Binomial{Float64},ProbitLink},GLM.DensePre
dChol{Float64,LinearAlgebra.Cholesky{Float64,Array{Float64,2}}}},Array{Float64,2}}

M ~ 1 + radius_mean + texture_mean + perimeter_mean + area_mean + smoothness_mean + compactness_mean + concavity_mean + symmetr
y_mean + fractal_dimension_mean + radius_se + texture_se + perimeter_se + area_se + smoothness_se + compactness_se + concavity_
se + symmetry_se + fractal_dimension_se + radius_worst + texture_worst + perimeter_worst + area_worst + smoothness_worst + comp
actness_worst + concavity_worst + symmetry_worst + fractal_dimension_worst

Coefficients:
────────────────────────────────────────────────────────────────────────────────────────────────
                              Coef.  Std. Error      z  Pr(>|z|)     Lower 95%    Upper 95%
────────────────────────────────────────────────────────────────────────────────────────────────
(Intercept)              -45.0247    27.5543     -1.63    0.1023   -99.0302       8.98086
radius_mean               -4.06648    9.69821    -0.42    0.6750   -23.0746      14.9417
texture_mean              -0.097076   0.186716   -0.52    0.6031    -0.463032     0.26888
perimeter_mean             0.432505   1.37841     0.31    0.7537    -2.26913      3.13414
area_mean                  0.0170119  0.0409039   0.42    0.6775    -0.0631582    0.097182
smoothness_mean          130.958     95.9243      1.37    0.1722   -57.0498     318.966
compactness_mean         -58.8173    49.8563     -1.18    0.2381  -156.534       38.8993
concavity_mean            64.3994    47.8947      1.34    0.1788   -29.4724     158.271
symmetry_mean              7.00524   24.6592      0.28    0.7763   -41.3259      55.3364
fractal_dimension_mean    60.315    188.651       0.32    0.7492  -309.435      430.065
radius_se                  9.08456   18.8901      0.48    0.6306   -27.9393      46.1085
texture_se                -2.33735    1.50642    -1.55    0.1208    -5.28988      0.615174
perimeter_se              -0.380159   1.8972     -0.20    0.8412    -4.09861      3.33829
area_se                    0.171095   0.171321    1.00    0.3179    -0.164688     0.506877
smoothness_se            606.774    297.509       2.04    0.0414    23.6663    1189.88
compactness_se          -110.29     128.398      -0.86    0.3904  -361.946      141.365
concavity_se             -41.0486    44.2424     -0.93    0.3535  -127.762       45.6649
symmetry_se             -154.209    131.734      -1.17    0.2418  -412.404      103.985
fractal_dimension_se   -1122.17     920.06       -1.22    0.2226 -2925.46       681.109
radius_worst               1.09282    3.10128     0.35    0.7246    -4.98557      7.17121
texture_worst              0.47468    0.230834    2.06    0.0397     0.0222535    0.927107
perimeter_worst            0.179667   0.33881     0.53    0.5959    -0.484388     0.843722
area_worst                -0.0215552  0.0349716  -0.62    0.5377    -0.0900982    0.0469879
smoothness_worst         -86.8069    59.8711     -1.45    0.1471  -204.152       30.5384
compactness_worst         -7.77587   23.906      -0.33    0.7450   -54.6307      39.079
concavity_worst           12.1558    17.0123      0.71    0.4749   -21.1877      45.4992
symmetry_worst            15.8367    14.835       1.07    0.2857   -13.2394      44.9128
fractal_dimension_worst  153.716    141.548       1.09    0.2775  -123.712      431.145
────────────────────────────────────────────────────────────────────────────────────────────────
```

In [ ]:

Fig: Model training

4.2 Accuracy of Model

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.



```
In [22]: # Accuracy of Logistic Regression Model
         # For Test Dataset
         accuracy = mean(prediction_df.correctly_classified)

Out[22]: 0.9847328244274809

In [ ]:
```

Fig: Accuracy of model on test dataset

5.0 Evaluating The Model

5.1 Confusion Matrix

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. **The confusion matrix shows the ways in which your classification model is confused when it makes predictions.** There could be four possible outcomes. Let us look at all four. -----**True Positives (TP)** - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.
True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.
False positives and false negatives, these values occur when your actual class contradicts with the predicted class.

False Positives (FP) – When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.
False Negatives (FN) – When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.
Once you understand these four parameters then we can calculate Accuracy, Precision, Recall and F1 score.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high

accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.803 which means our model is approx. 80% accurate.
Accuracy = TP+TN/TP+FP+FN+TN

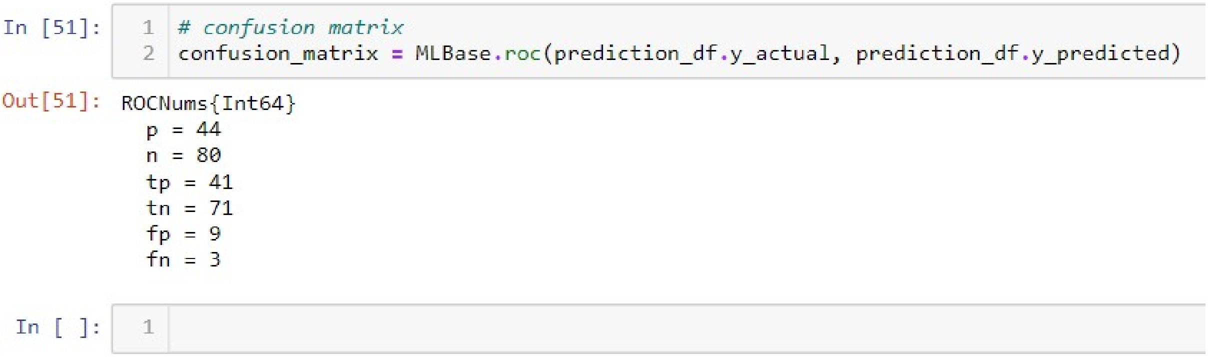
Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labelled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.
Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: Of all the passengers that truly survived, how many did we label? We have got recall of 0.631 which is good for this model as it's above 0.5.
Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, F1 score is 0.701.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

The support is the number of samples of the true response that lie in that class.



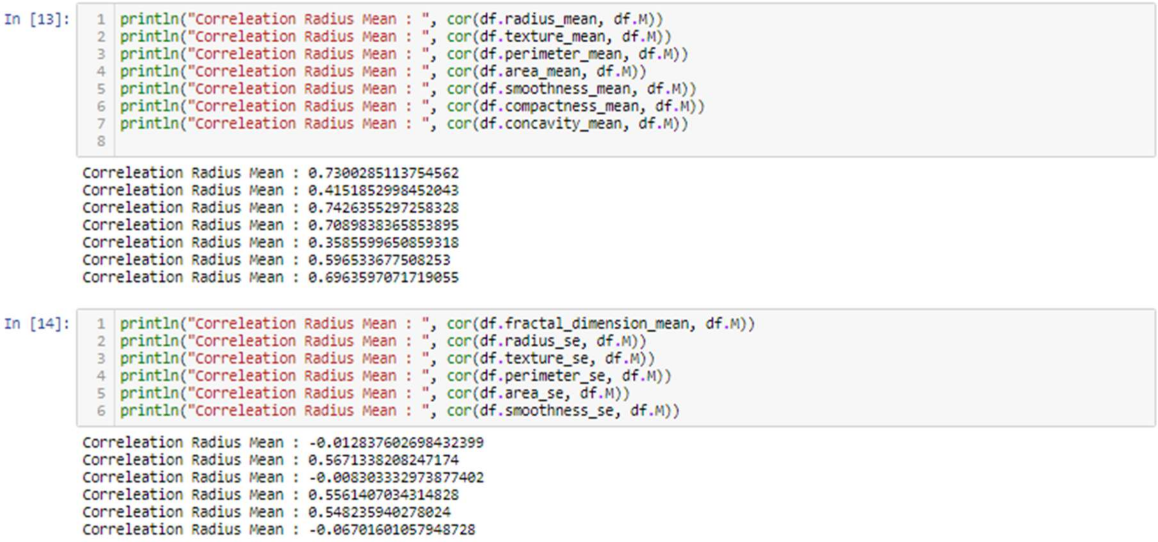
```
In [51]: # confusion matrix
         confusion_matrix = MLBase.roc(prediction_df.y_actual, prediction_df.y_predicted)

Out[51]: ROCNums{Int64}
           p = 44
           n = 80
           tp = 41
           tn = 71
           fp = 9
           fn = 3

 In [ ]:
```

Fig: Confusion Matrix output

6.0 Improving Model

6.1 To improve the accuracy of model, following steps can be followed –

1. **Feature Scaling and/or Normalization** - Check the scales of your *gre* and *gpa* features. They differ on 2 orders of magnitude. Therefore, your *gre* feature will end up dominating the others in a classifier like Logistic Regression. You can normalize all your features to the same scale before putting them in a machine learning model.
2. **Class Imbalance** - Look for class imbalance in data. Since you are working with admit/reject data, then the number of rejects would be significantly higher than the admits. Most classifiers in such as Logistic Regression have a `class weight` parameter. Setting that to `balanced` might also work well in case of a class imbalance.
3. **Optimize other scores** - You can optimize on other metrics also such as *Log Loss* and *F1-Score*. The F1-Score could be useful, in case of class imbalance.
4. **Hyperparameter Tuning - Grid Search** - You can improve your accuracy by performing a Grid Search to tune the hyperparameters of your model. For example, in case of `LogisticRegression`, the parameter `C` is a hyperparameter. Also, you should avoid using the test data during grid search. Instead perform cross validation. Use your test data only to report the final numbers for your final model. Please note that GridSearch should be done for all models that you try because then only you will be able to tell what the best is you can get from each model.
5. **Correlation** - If we add a lot of correlated features in our model, we may cause the model to consider unnecessary features and we may have curse of high dimensionality problem.



```
In [13]:
println("Correleation Radius Mean : ", cor(df.radius_mean, df.M))
println("Correleation Radius Mean : ", cor(df.texture_mean, df.M))
println("Correleation Radius Mean : ", cor(df.perimeter_mean, df.M))
println("Correleation Radius Mean : ", cor(df.area_mean, df.M))
println("Correleation Radius Mean : ", cor(df.smoothness_mean, df.M))
println("Correleation Radius Mean : ", cor(df.compactness_mean, df.M))
println("Correleation Radius Mean : ", cor(df.concavity_mean, df.M))
```

```
Correleation Radius Mean : 0.7300285113754562
Correleation Radius Mean : 0.4151852998452043
Correleation Radius Mean : 0.7426355297258328
Correleation Radius Mean : 0.7089838365853895
Correleation Radius Mean : 0.3585599650859318
Correleation Radius Mean : 0.596533677508253
Correleation Radius Mean : 0.6963597071719055
```

```
In [14]:
println("Correleation Radius Mean : ", cor(df.fractal_dimension_mean, df.M))
println("Correleation Radius Mean : ", cor(df.radius_se, df.M))
println("Correleation Radius Mean : ", cor(df.texture_se, df.M))
println("Correleation Radius Mean : ", cor(df.perimeter_se, df.M))
println("Correleation Radius Mean : ", cor(df.area_se, df.M))
println("Correleation Radius Mean : ", cor(df.smoothness_se, df.M))
```

```
Correleation Radius Mean : -0.012837602698432399
Correleation Radius Mean : 0.5671338208247174
Correleation Radius Mean : -0.008303332973877402
Correleation Radius Mean : 0.5561407034314828
Correleation Radius Mean : 0.548235940278024
Correleation Radius Mean : -0.06701601057948728
```

Fig: Correlation among different features

7.0 Conclusion

In Germany the quality of oncological care is already very high. This is also due, in no small measure, to the establishment of certified facilities for oncological care. In addition to the positive effect of improving the quality indicators (QI) which serve as indirect parameters for the quality of outcomes certified centers have had a highly significant impact on improving the mortality of breast cancer patients – as demonstrated by the 3 certified centers in Middle Franconia. Similar results have also been reported for a single-center study carried out in the Breast Center of the University Hospital Heidelberg. But care providers are currently burdened with additional costs for which they are not being adequately reimbursed. The financial and staffing costs required to document breast cancer patients are considerable; in addition to clinical data, considerable time and resources are spent on documentations for quality assurance and quality management required as part of certification, notwithstanding the fact that clinical documentation can also be very detailed, as these data are used secondarily for quality assurance. In the long term, the current high standard of care can only be maintained if there is adequate financial support and if facilities are relieved of some of the costs of documentation. The most important aspect of this expenditure is the financial and staffing cost currently needed for documentation. Measures to reduce the time and cost of documentation are urgently required. A more effective way of collecting and collating data is necessary. Investment in suitable documentation systems with compatible interfaces is necessary. The subsequent impact on healthcare could be considerable:

1. Staffing and financial costs currently used for documentation could be reduced and the resources could be invested in other areas of the healthcare system.
2. This could relieve some of the burden on doctors, many of whom are already working at the limits of their capacity, particularly as there is an increasing shortage of young people entering the profession – even if enough money were available to leave documentation primarily in the hands of doctors, this would not be possible due to the lack of doctors.
3. This could strengthen the position of new professional groups working in healthcare, for example medical documentation assistants specializing in tumour documentation.
4. Quality assurance could be optimized by defining fewer but more relevant quality indicators and ensuring that these data are documented in the same standard format by all professional groups and medical specialties.

Based on the findings of this single-center project, a multi-center survey will be carried out to validate the results of this study and to highlight differences in documentation costs and times in facilities offering different levels of care as well as differences between certified and non-certified facilities.

8.0 Future Scope of Project

Breast cancer remains the most common invasive cancer among women. The primary patients of breast cancer are adult women who are approaching or have reached menopause; 90 percent of new cases in U.S. women in 2009 were diagnosed at age 45 or older. Growing knowledge of the complexity of breast cancer stimulated a transition in breast cancer research toward elucidating how external factors may influence the etiology of breast cancer.

Breast Cancer and the Environment reviews the current evidence on a selection of environmental risk factors for breast cancer, considers gene-environment interactions in breast cancer, and explores evidence-based actions that might reduce the risk of breast cancer. The book also recommends further integrative research into the elements of the biology of breast development and carcinogenesis, including the influence of exposure to a variety of environmental factors during potential windows of susceptibility during the full life course, potential interventions to reduce risk, and better tools for assessing the carcinogenicity of environmental factors. For a limited set of risk factors, evidence suggests that action can be taken in ways that may reduce risk for breast cancer for many women: avoiding unnecessary medical radiation throughout life, avoiding the use of some forms of postmenopausal hormone therapy, avoiding smoking, limiting alcohol consumption, increasing physical activity, and minimizing weight gain.

Breast Cancer and the Environment sets a direction and a focus for future research efforts. The book will be of special interest to medical researchers, patient advocacy groups, and public health professionals.

9.0 References

1. Robert Koch-Institut ; Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V., Hrsg . Berlin: Robert Koch-Institut; 2013. Krebs in Deutschland 2009/2010. 9. Ausgabe. [Google Scholar]

2. Bundesministerium für Bildung und Forschung Bevölkerung in Deutschland nach Alter und Geschlecht (Tabelle 0.14). 2012Online:http://www.datenportal.bmbf.delast access: 08.05.2014

3. Lüftner D, Lux M P, Maass N. et al.Advances in breast cancer – looking back over the year. Geburtsh Frauenheilk. 2012;72:1117–1129. [PMC free article] [PubMed] [Google Scholar]

4. Beckmann M W. Nationaler Krebsplan des Bundesministeriums für Gesundheit – Strategieplan der Deutschen Krebsgesellschaft. Frauenheilkunde up2date. 2009;5:323–329. [Google Scholar]

5. Bundesministerium für Gesundheit Nationaler KrebsplanOnline:http://www.bmg.bund.de/cln_169/nn_1168248/SharedDocs/Standardartikel/DE/AZ/N/Glossarbegriff-Nationaler-Krebsplan.html%23doc1632822bodyText14last access: Februar 2010

6. Kowalski C, Wesselmann S, Kreienberg R. et al.The patients' view on accrediated breast cancer centers: Strengths and potential for improvement. Geburtsh Frauenheilk. 2012;72:137–143. [PMC free article] [PubMed] [Google Scholar]

7. Querschnitts-AG Dokumentation Datensparsame einheitliche Tumordokumentation – eine Kernforderung des Nationalen Krebsplans. Endfassung 10.10.2011Online:http://www.tumorzentren.de/Nationaler-Krebsplan.htmllast access: 07.08.2014

8. Tarifgemeinschaft deutscher Länder, Hrsg. Tarifvertrag für Ärztinnen und Ärzte an Universitätskliniken (TV-Ärzte). vom 30. Oktober 2006 in der Fassung des Änderungstarifvertrages Nr. 1 vom 27. August 2009Online:http://www.uni-erlangen.de/einrichtungen/personalabteilung/last access: 17.04.2011

9. Bundesärztekammer Hrsg.Gehaltstarifvertrag für Medizinische Fachangestellte. In: Deutsches Ärzteblatt (106/24)Online:http://www.aerzteblatt.de/download/files/2009/06/down137850.pdflast access: 24.11.2011

10. Kolberg H-C, Lüftner D, Lux M P. et al.Breast cancer 2012 – New aspects. Geburtsh Frauenheilk. 2012;72:602–615. [PMC free article] [PubMed] [Google Scholar]

11. Kreienberg R, Albert U-S, Follmann M. et al.Interdisciplinary GoR level III guidelines for the diagnosis, therapy and follow-up care of breast cancer. Geburtsh Frauenheilk. 2013;73:556–583. [PMC free article] [PubMed] [Google Scholar]

12. Brucker S Y, Schumacher C, Sohn C. et al.Benchmarking the quality of breast cancer care in a nationwide voluntary system: the first five-year results (2003–2007) from Germany as a proof of concept. BMC Cancer. 2008;8:358. [PMC free article] [PubMed] [Google Scholar]

13. Brucker S Y, Bamberg M, Jonat W. et al.Certification of breast centres in Germany: proof of concept for a prototypical example of quality assurance in multidisciplinary cancer care. BMC Cancer. 2009;9:228. [PMC free article] [PubMed] [Google Scholar]

14. Beckmann M W, Brucker C, Hanf V. et al.Quality assured health care in certified breast centers and improvement of prognosis of breast cancer patients. Onkologie. 2011;34:362–367. [PubMed] [Google Scholar]

15. Heil J, Gondos A, Rauch G. et al.Outcome analysis of patients with primary breast cancer initially treated at a certified academic breast unit. Breast. 2012;21:303–308. [PubMed] [Google Scholar]

16. Beckmann M W, Bani M R, Loehberg C R. et al.Are certified breast centers cost-effective? Breast Care. 2009;4:245–250. [PMC free article] [PubMed] [Google Scholar]

17. Lux M P, Hildebrandt T, Beyer-Finkler E. et al.Relevance of health economics in breast cancer treatment – the view of certified breast centres and their patients. Breast Care. 2013;8:15–21. [PMC free article] [PubMed] [Google Scholar]