Title of Project

Survival Analysis (Human Breast Cancer Prediction)

### Insightful Coders Team

**Team Members Names**

* **Divi Niveditha Nirmala**
* **Patchipulusu Vaishnavi**
* **Apoorva Sritha Reddy P**
* **Divya Raman**
* **Ravela Durga Laasya**

**Content Page**

**1. Title of the project**

**1.1 Introduction**

**1.2 Objectives of research**

**1.3 Problem Statement**

**1.4 Industry Profile**

**2. Review of literature**

**3. Data Collection**

**4. Methodology**

**4.1 Exploratory Data Analysis**

**4.1.1 Figures and tables**

**4.2 Statistical Techniques and Data Visualization**

**4.3 Data Modeling using Supervised ML techniques**

**5. Findings and Suggestion**

**6. Conclusion**

**7. References**

**Introduction**

**Cancer**, also called “**malignancy”**, is an abnormal growth of cells. There are more than 100 types of cancer, including breast cancer, skin cancer, lung cancer etc. Symptoms vary depending on the type.

**Breast cancer** is the most common type of cancer for women regardless of race and ethnicity. Cancer treatment may include chemotherapy, radiation, and/or surgery.

**Some Risk Factors for Breast Cancer**

The following are some of the known risk factors for breast cancer. However, most cases of breast cancer cannot be linked to a specific cause. Talk to your doctor about your specific risk.

**Age:**  The chance of getting breast cancer increases as women age. Nearly 80 percent of breast cancers are found in women over the age of 50.

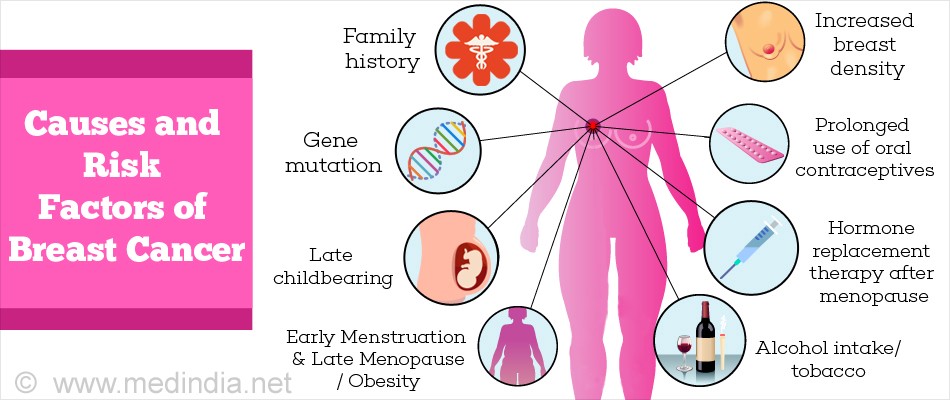
**Personal history of breast cancer:**  A woman who has had breast cancer in one breast is increased the risk of developing cancer in her other breast.

**Family history of breast cancer:**  A woman has a higher risk of getting breast cancer. If her mother, sister or daughter had breast cancer, especially at a young age (before 40), there may be a high chance of getting breast cancer to that lady.

**Genetic factors:**  Women with certain genetic mutations, including changes to the BRCA1 and BRCA2 genes, are at higher risk of developing breast cancer during their lifetime. Other gene changes may raise breast cancer risk as well.

**Childbearing and menstrual history:** The older a woman is i.e., when she has her first child, the greater her risk of breast cancer. Also at higher risk are:

* Women who menstruate for the first time at an early age (before 12)
* Women who go through menopause late (after age 55)
* Women who’ve never had children



**Objectives of Research**

Breast cancer is the most common cancer among women and one of the major causes of death among women worldwide. Every year approximately 124 out of 100,000 women are diagnosed with breast cancer, and the estimation is that 23 out of the 124 women will die of this disease.

When detected in its early stages, there is a 30% chance that the cancer can be treated effectively, but the late detection of advanced-stage tumors makes the treatment more difficult. Currently, the most used techniques to detect breast cancer in early stages are:

* Mammography (63% to 97% correctness)
* FNA (Fine Needle Aspiration**)** with visual interpretation (65% to 98% correctness)
* surgical biopsy(approximately 100% correctness).

Therefore, mammography and FNA with visual interpretation correctness varies widely, and the surgical biopsy, although reliable, is invasive and costly.

Step Wise Procedure

🡺Building a classifier using machine learning can be a difficult task if the dataset used is not on its best format or if it is not being correctly interpreted. Therefore, a considerable portion of this work will be spent preparing and comprehending the dataset in order to avoid problems such as overfitting. To prepare the dataset, we used Preprocessing techniques and prepare the training set before it can generate the classifier.

Some machine learning techniques are compared here.

Methods are used to create classiffiers that must discriminate benign from malignant breast lumps. To create the classiffier, the “**WBCD”** (Wisconsin Breast Cancer Diagnosis) dataset is employed. This dataset is widely utilized for this kind of application because it has a large number of instances (569), is virtually noise-free and has no missing values. Before performing the tests, a large fraction of this work will be dedicated for pre-processing the data in order to optimize the classiffier.

🡺The next step is to propose methods and algorithms to optimize the training set.

* Are there any outliers?
* How many malignant and benign are there in the data set?
* What is the correlation between independent variable and target variable?
* What is the correlation between the independent variable?

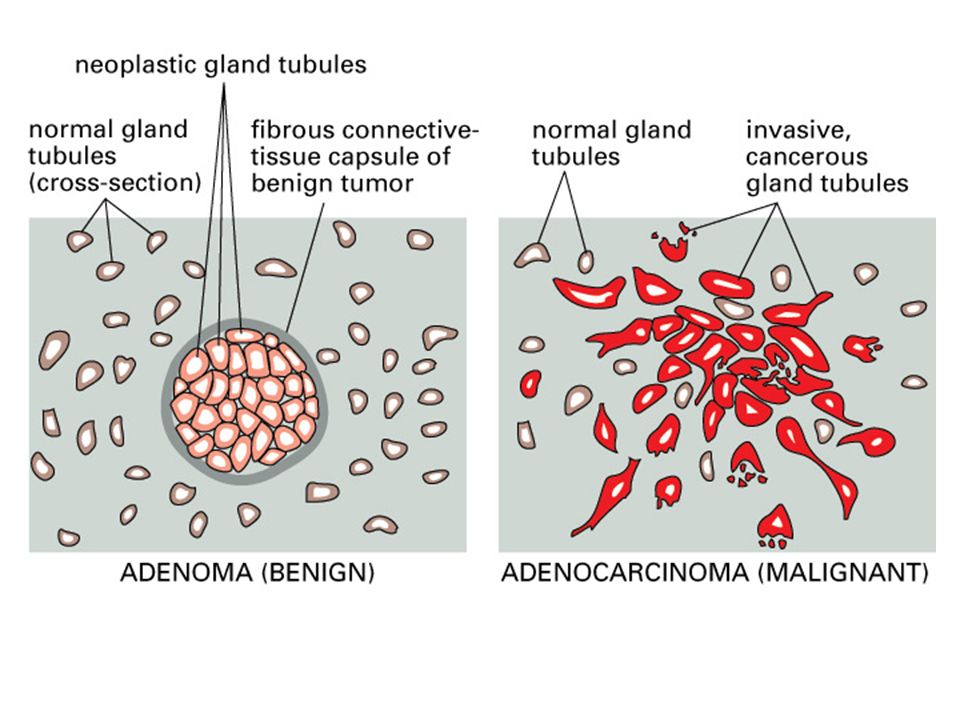
All these questions are discussed and different solutions are proposed.

The results are presented in tables, which contains the accuracy of the classiffer, the rate of false-negatives and the rate of false-positives. All the tests were conducted in Jupiter Notebook and machine learning techniques like pre-processing, scaling, classiffication were also performed.

**Problem Statement**

The problem statement deals with the determination of cell i.e., malignant or benign. Based on the cell parameters such as radius\_mean, texture\_mean etc… the model predicts the type of tumor (Benign or Malignant). Malignant is cancerous, while benign is non-cancerous. We predict this with the help of cell parameters.

This analysis is helpful in determining if a cell is cancerous or not in an early stage. This will help in saving the precious human life.



**Industry profile**

The health care industry is tremendously important to people around the world. This industry comprises of different players including hospitals, doctors, nursing homes, diagnostic laboratories, pharmacies, medical device manufacturers, and other components of the health care system.

The health care industry, or medical industry, is a sector that provides goods and services to treat patients with curative, preventive, rehabilitative or palliative care. The healthcare industry is composed of establishments devoted to prevention, diagnosis, treatment, and rehabilitation of medical conditions. Such treatment may be through providing products or services, and may be provided privately or publicly. The modern health care sector is divided into many sub-sectors, and depends on interdisciplinary teams of trained professionals and paraprofessionals to meet health needs of individuals and populations. The health care industry includes establishments ranging from small-town private practices of physicians who employ only one medical assistant to busy inner-city hospitals that provide thousands of diverse jobs. Healthcare industry is littered with risks and challenges as it is an industry that requires constant innovation under increased regulations.

This model is of extremely useful to the health industry especially the cancer institutes. This will help them make predictions about the cell with the help of parameters so that subjective human errors can be reduced.

**Review of literature**

There has been a lot of research on medical diagnosis of breast cancer. Many of those consider models that can predict whether the tumour is benign or malignant. For predicting such a classification problem many techniques are available. Besides this, different kinds of explanatory variables can be used for predicting the presence of malignant tumour cells. A lot of researches, however, have used one of the three different Wisconsin breast cancer datasets (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC), and Wisconsin Prognosis Breast Cancer (WPBC)). All three datasets are used in the paper of Salama et al. [16], where each dataset has different features trying to predict the outcome. For each of the datasets, a multi-classifier was made using a selection of the classifiers decision tree, Multi-Layer Perceptron, Naive Bayes, Sequential Minimal Optimization, and Instance Based for K-Nearest neighbour. **The highest obtained accuracies** **are 97.28%, 97.72%, and 77.32% for the datasets WBC, WDBC, and** **WPBC**, respectively. Several studies used the WBC dataset for detecting breast cancer. In the study of Karabatak et al. [11] breast cancer is detected based on association rules and a neural network. The association rules are used for eliminating unnecessary data and thus reducing the feature dimension. The neural network classifies each record using those remaining features. The final model resulted in an accuracy of 97.4%. Abbass [1] achieved an accuracy of 98.1% ± 0.5 using an evolutionary artificial neural network approach. This approach is based on the Pareto-differential evolution algorithm that is augmented with local search. In Marcano-Cedeño et al. [14], an Artificial metaplasticity Multilayer Perceptron algorithm is applied, obtaining a classification accuracy of 99.26%. Akay [2] reached the highest accuracy of 99.51% using an SVM model combined with feature selection. Various other studies used the WDBC dataset, which is also used in this paper. Wolberg et al. [18] applied two models, logistic regression and Multisurface Method-Tree. These resulted in 10-fold cross-validated classification accuracies of 96.2% and 97.5%, respectively. In the research of Mu et al. [15] support vector machines, radial basis function networks, and self-organizing maps are applied to detect breast cancer. The performance of different combinations of the classifiers is compared based on 10-fold cross-validation. The average performance accuracy is over 98%.

**Data collection**

The dataset used in this paper is publically available[8] and

was created by Dr. William H. Wolberg, physician at the

University Of Wisconsin Hospital at Madison, Wisconsin,

USA. It was donated by Olvi Mangasarian on July 15th,

1992 [9].

To create the dataset Dr. Wolberg used ﬂuid samples, taken

from patients with solid breast masses[10] and an easy-to-

use graphical computer program called Xcyt[11], which is

capable of perform the analysis of cytological features based on a digital scan.

The program uses a curve-ﬁtting algorithm to compute ten features from each one of the cells in the sample, than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector. Each feature is evaluated on a scale of 1 to 10, with 1 being

the closest to benign and 10 the closest to malignant. Statistical analysis showed that the following nine characteristics diﬀer signiﬁcantly between benign and malignant samples: uniformity of cell shape, uniformity of cell size, clump thickness, bare nuclei, cell size, normal nucleoli, clump cohesiveness, nuclear chromatin and mitoses.

The samples were taken periodically as Dr. Wolberg reported his clinical cases; therefore the data is presented as chronological groups that reﬂect the period they were created.

* **Dataset started being built (January 1989).**
* **Until the last instance created (November 1991)**
* **Class distribution: 357 benign, 212 malignant**

**Source:**

1. Dr. William H. Wolberg, General Surgery Dept.

University of Wisconsin, Clinical Sciences Center Madison, WI 53792

wolberg '@' eagle.surgery.wisc.edu

1. W. Nick Street, Computer Sciences Dept.

University of Wisconsin, 1210 West Dayton St., Madison,

WI 53706

street '@' cs.wisc.edu 608-262-6619

3. Olvi L. Mangasarian, Computer Sciences Dept.

University of Wisconsin, 1210 West Dayton St., Madison, WI 53706

olvi '@' cs.wisc.edu

**Attribute Information:**

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32)

Ten real-valued features are computed for each cell nucleus:

* **Radius:** The average distance from the center of the nucleus to each of the boundary points.
* **Texture:** The standard deviation of the gray-scale values. A gray-scale value represents the intensity of the shades of gray in each pixel of the image.
* **Perimeter:** The total distance of the boundary of the cell nucleus.
* **Area:** The number of pixels on the interior of the boundary and adding one-half of the pixels on the perimeter, to correct for the error caused by digitization.
* **Smoothness**: The difference between the length of a radius length and the mean length of the two radius lines surrounding it, hence the local variation in radius lengths.
* **Compactness**: The perimeter and area are combined to obtain a measure of compactness of the cell nuclei.
* **Concavity**: The severity of concave portions of the contour. A high concavity means that the boundary of the cell nucleus has indentations, and thus is rather rough than smooth.
* **Concave** **points**: The number of concave portions of the contour of the cell nucleus.
* **Symmetry**: The symmetry is determined by first finding the longest line from boundary point to boundary point through the center of the nucleus. Subsequently, the relative length differences between the lines perpendicular to the longest line to the boundary in both directions are measured. Attention should be given to nuclei where the longest line cuts through the boundary because of concavity
* **Fractal** **dimension**: The fractal dimension is approximated by the ’coastline approximation’. The perimeter of the nucleus can be measured using different lengths of measuring sticks. As this length increases, the total length of the measured ’**coastline’** decreases due to lower precision of the 8 measurement.

**Methodology**

**Exploratory Data Analysis**

In the dataset, we find that there are 31 columns (i.e., independent variables) and 569 rows. Based on the correlation between the target variable and the independent variables we drop the columns by taking the columns which are having very less (nearer to 0) correlation and also the variables which are redundant(based on the heat maps).

Here for this model, the independent variables are

* radius\_mean
* texture\_mean
* smoothness\_mean
* compactness\_mean
* concavity\_mean
* symmetry\_mean
* radius\_se
* compactness\_se
* concavity\_se
* concave\_points\_se
* radius\_worst
* texture\_worst
* smoothness\_worst
* compactness\_worst
* concavity\_worst
* concave\_points\_worst
* symmetry\_worst
* fractal\_dimension\_worst.

The dependent variable is ‘**diagnosis’**.

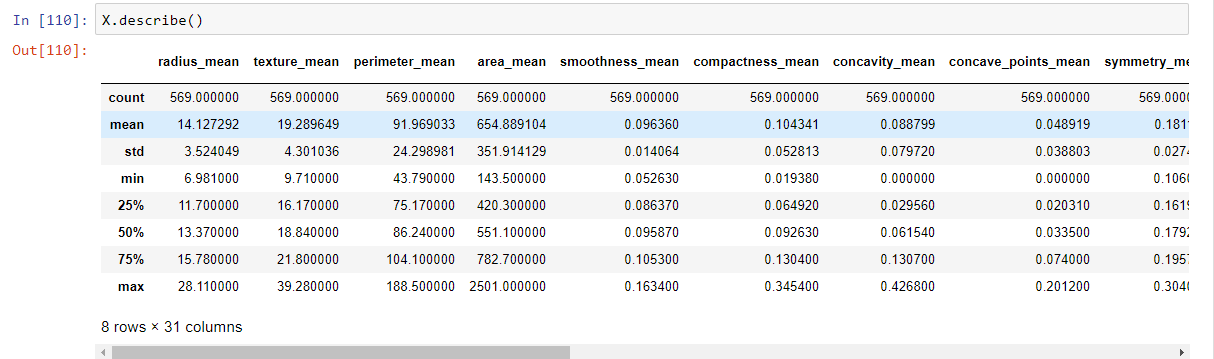
We don’t consider the column 'id' in analysis because this column is not needed for training the model.

**FIGURES AND TABLES**:

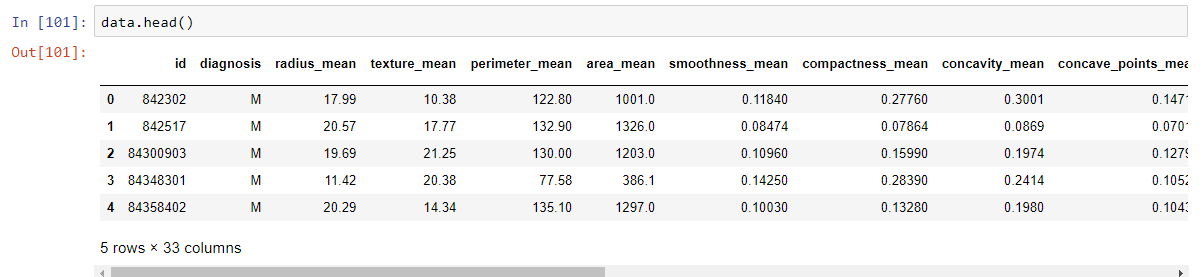
* **Data Information**:



* **Data Description**:



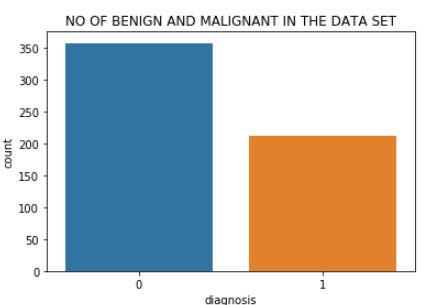
* **Data Head**:



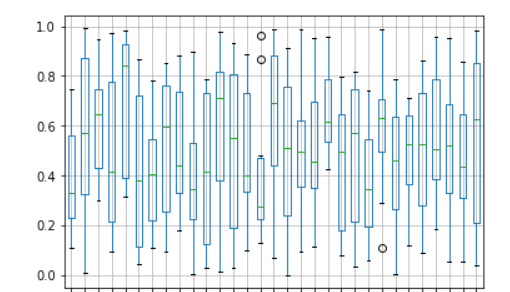
* In ‘**diagnosis’** column,

Number of Benign: 357

Number of Malignant: 212



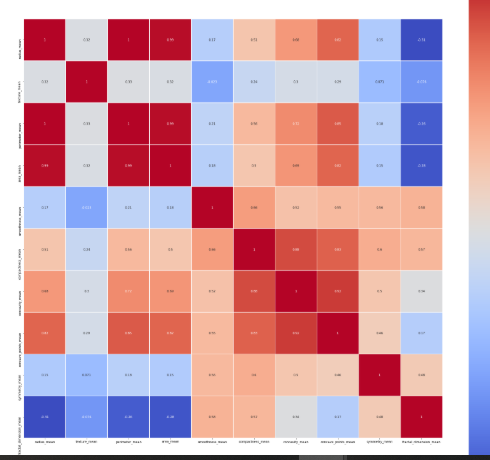
On plotting and analyzing the **Box plot** , we understand that there are very less outliers in the columns.



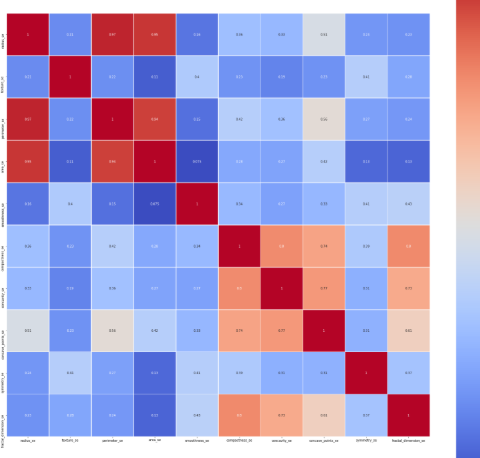
Since it is a classification problem, we apply various algorithms, and find the algorithm that has **higher ROC\_SCORE**.

**STATISTICAL TECHNIQUES AND DATA VISUALIZATION**

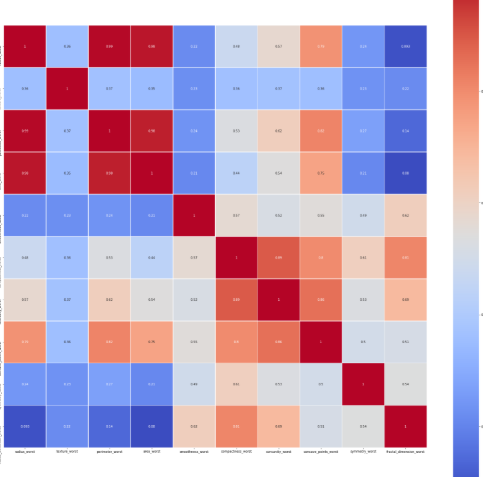
* **Heat Map for mean variables**



* **Heat Map for Standard Error**



* **Heat Map for Worst**



**DATA MODELLING USING SUPERVISED ML TECHNIQUES**

***Logistic Regression****:*

**“Regression** is a statistical process that estimates the relationship between the dependent (target) variable and the independent variables”. **Linear regression** is the most basic type of regression, in which case the relation between the dependent and independent variable(s) is linear.

**Logistic regression** models can be thought of as extensions of linear regression models, and hence, linear regression is explained first. In general, there are n observed data points

y1, ..., yn, which represent the realizations of the independent random variables Y1, ..., Yn. Besides this, there are m explanatory variables for which the right coefficients vector β needs to be found. The vector xi of length m+1 contains the intercept along with the explanatory variables of the i-th observation.

**The linear regression model can be described as follows**.

1. Yi ∼ N (µi , σ2 ),
2. ηi = x T i β,
3. ηi = g(µi) = µi ,

for i = 1,...,n, with β = (β0, ..., βm) T

The coefficient vector for the intercept (β0) and the explanatory variables. This representation seems quite difficult for such a relatively easy model, but it ensures that the alteration to the logistic regression model is clear. As shown in above, the model consists of three components, the random component

* (i)The systematic component
* (ii)The link function
* (iii)The random component specifies the distribution of the target Yi .

The systematic component is the vector ηi , which consists of the predictors for each observation. These predictors are formed by multiplying the values of the explanatory variables with the coefficient vector. The “*link function”,* denoted by g, is the link between the random and the systematic component. More precisely, it specifies the relation between the two by **g(EYi) = ηi .** Since the relation is linear ηi equals µi . In logistic regression (LR), however, the dependent variable is categorical. The approach of logistic regression is part of a whole class of models, called **generalized linear models (GLMs).** They are called generalized linear models because the systematic component remains the same, and hence the ηi are still linear. The random component and the link function need to be adjusted, though, to 13 transform the linear regression model into the logistic regression model. In the linear regression model ηi and Yi can take any real number, meaning ηi , Yi ∈ R. In logistic regression the target variables Yi can only take two values 0 or 1. This means the following for the binomial random variables, 0 < EYi < 1. And, therefore, the link function should ensure that it maps the interval (0,1) onto the real line. Satisfying this condition results in the logistic regression model below.

1. niYi ∼ Bin(ni , µi),
2. ηi = x T i β,
3. ηi = g(µi) = log µi 1−µi ,

for i = 1,...,n. Relating the theory to this particular problem, the binary dependent variable can take one of the two values **’benign’** or **’malignant’**. Then the algorithm tries to fit the best regression coefficient vector β to the given data.

**Using logistic regression, we got the classification report as follows**

**precision recall f1-score support**

**0 0.99 0.97 0.98 108**

**1 0.95 0.98 0.97 63**

**micro avg 0.98 0.98 0.98 171**

**macro avg 0.97 0.98 0.98 171**

**weighted avg 0.98 0.98 0.98 171**

**The confusion matrix**:

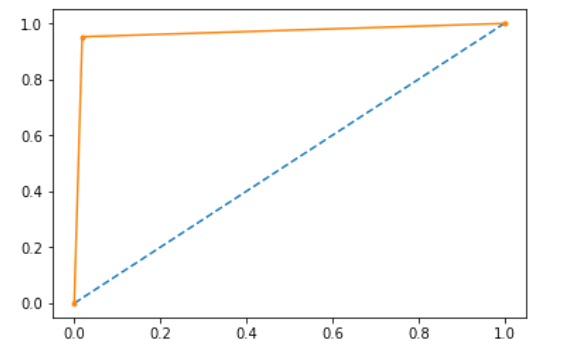
[[105 3]

[ 1 62]]

**We find that there are 3 false positives and 1 false negative.**

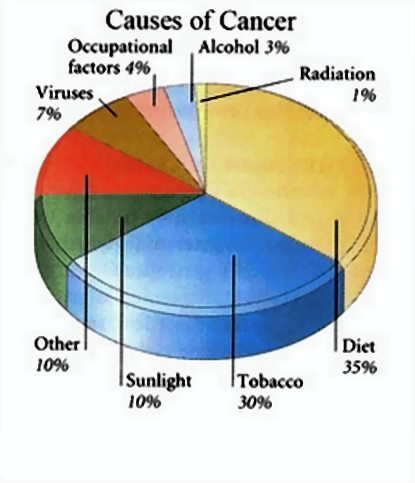
**We get the ROC SCORE as 0.9781746031746033.**

\***roc curve**\*



**Findings and suggestions:**

We found out that out of 569 cases , 212 cases are malignant cells, which forms 37.2% of the total cases reported. It shows that less than 50% of the cases are cancerous, which can further be reduced by taking appropriate measures and taking care the health and consuming nutritrous food. It is found that 35% of cancer is caused by diet which forms the highest percent.



**Conclusions:**

From the results it can be concluded that all five models obtain very promising performances in classifying the possible breast cancer. All models are optimized based on the accuracy, hence the final model should at least have the highest accuracy. Selecting the best-suited model for this specific problem also depends on the sensitivity value, because it is important to have a low number of false positives. The tumour cell nuclei are best predicted by the support vector machine and the ensemble. Both have the highest performance values for accuracy, sensitivity, and specificity. However, the support vector machine is the model which also has the highest value for the AUC. Therefore, the support vector machine model is recommended to use for this specific problem.

Since all the models easily have performance values over 90%, it can be concluded that the features have a high predictive power. This not only might be a reason why all models have such high but also significantly similar performances. Thus, during further research into this problem, all five models are suitable to optimize towards extremely high performance values.

**Based on ROC SCORE of various classification algorithms, we chose Logistic Regression as our model becauseit has got highest score with 97.81%.**

**References**:

W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

<http://rexa.info/paper/b98475235164960529ad2ff9fda3816e9335cf8a>

O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.

<http://rexa.info/paper/90e988e83c7f06d2797b41580569c1f9a13f6749>

Medical literature:   
  
W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171. 

\*\*\* THANKING YOU \*\*\*