

BREAST CANCER WISCONSINPROGNOSTIC

Name: Sonu Sankhala

Roll No.: 12

Registration No. 11805066

Section: KM007

# ABOUT DATASET

This dataset is consist 12 features and each feature is containing different values like **radius, smoothness,** etc. They describe characteristics of the cell nuclei present in the image. It also shows certain parameter through which we can classify whether a person is having breast cancer or not. This dataset is having 569 instances. It is having 32 attributes.

**Features of dataset are as follows:**

1. ID number
2. Diagnosis(M= malignant, B=bengin)

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

1. Radius(mean of distances from center to points on the perimeter)
2. Texture(standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness(local variation in radius lengths)
6. Compactness(perimeter^2/area-1.0)
7. Concavity(severity of concave portions of the contour)
8. Concave points(number of concave portions of the contour)
9. Symmetry
10. Fractal dimension(“coastline approximation”-1)

Feature Visualization

A.) Decision tree

Timeline

Description automatically generated

B.)Accuracy VS Number of Tree

A picture containing line chart

Description automatically generated

C.) Malignant & Bengin

A picture containing shape

Description automatically generated

Algorithm and Techniques Used and Accuracy Achieved

**Random Forest Classifier:** Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees**.**

I got Accuracy of 92% from Random Forest Classifier.

**Decision Tree:** A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels

I got Accuracy of 89% from Decision Tree Classifier.

**Logistic Regression:** Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

I got Accuracy of 93% from Logistic Regression Classifier.

**Technique Used for Training Model:**

**Scikit Learn:** Scikit-learn is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

**Accuracy achieved after training model:**

Highest accuracy on testing dataset is 93% and that come with the Logistic Regression model.

**Thank you**