**BREAST CANCER PREIDICTION**

**WHAT IS BREAST CANCER :** Breast cancer is a type of cancer that starts in the breast. It can start in one or both breasts.

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society.

This machine learning project uses a dataset that can help determine the likelihood that a breast tumor is malignant or begin .Various factors are taken into consideration, including the lump’s thickness, number of bare nuclei, and mitosis. This is also an excellent way for new machine learning professionals to practice R programming.

**ANALYSIS:**

* Frist import required librabries. Here, we required Pandas library so import pandas library.
* Later read csv file which is already uploaded like using code

df = pd.read\_csv(“Breast cancer prediction data.csv”)

* head() method returns first 5 rows of the given data.
* tail() method returns last 5 rows of the given data.
* describe() method returns description of the data in dataframe.
* type() method returns type of that dataframe.
* .Shape returns no.of rows and columns of dataframe.
* dtypes returns datatype of columns like int64or float64.
* isnull() returns true where that dataset contains not a null value else false.
* duplicate() method returns a Boolean series which is True only for Unique elements.
* The function dataframe.isnull().sum().sum() returns the no.of missing values in the dataset.
* cols=list(df.columns) returns list of columns mean columns names.
* Again import required libraries.
* sns.countplot, Seaborn literally counts the no.of observations per category for a categorical variables,and displays the results as a bar chart.
* Seaborn heatmap is a graphical representation of 2D data.
* PairGrid returns a grid of subplots using the same plot type to visualize data.
* Pandas MinMaxscalar transform feature by scaling each feature to given range.
* Iloc() function in python helps us to select a specific row or column from the dataset.
* Corr() function to find the pairwise correlation of all columns in the pandas DataFrame in python.
* Groupby() .mean () returns the mean of each type in the given column.
* Groupby().count() returns the count of each type in the given column.
* Fit(x\_train,y\_trian) evaluates how the classifier performs on the training set with.
* Score(x\_train,y\_train) is measuring the accuracy of the model against the training data.
* Spilt() is used to splits data into train and test to verify accuracy after fitting the model.

**Preprared by**

**Ch.geetha(s180402)**

**R.laxmisri(s181016)**