

National Institute of Technology, Jamshedpur

**COMPUTER SCIENCE AND ENGINEERING DEPARTMENT**

**Soft Computing Lab**

**PROJECT REPORT**

**ON**

**“Breast Cancer Detection”**

*Submitted by*

**Akhilesh Kumar Mishra 2020UGCS053**

**Angshuman Pal 2020UGCS105**

**Rajdeep Kumar 2020UGCS070**

**Ankur Kumar 2020UGCS037**

***Under the guidance of***

**Mr. Koushlendra Kumar Singh**

Sr. Assistant Professor,

Dept. of CSE

**ACKNOWLEDMENT**

We as team were very excited when we were assigned this project on **Breast- Cancer-Detection**. In the world when Artificial Intelligence along with Machine Learning has boundless capabilities to help in every filed, the medical field is one of the prominent among them.

Whatever we have able to create was far impossible without the mention of the people who mase it possible, whose constant guidance and encouragement crowned our efforts with success.

We have great pleasure in expressing my deep gratitude to **Dr. Koushelndra Kumar Singh**, Associative professor of Computer Science Department for assigning this project to us and showing us the wide usability of various Machine Learning Algorithms.

Along with the Professor In charge We would like to **Dr**. **Geeta Kadayaprath** Oncoplastic Breast Surgeon, Senior Director at Max institute of Cancer Care, Max Healthcare for taking the Workshop Kary Shala 2022 at our College that ignited interest of our team to work upon this project.

Finally, a note of thanks to the all my friends, teaching and non-teaching staff, Lab Assistants of Dept of Computer Science and Engineering, for their cooperation extended to us.

**ABSTRACT**

Breast Cancer Detection using machine learning comes into picture because the doctors do not identify each and every patient. That’s why the number of women affected by breast cancer are still increasing.

So, we are using some models for detecting breast cancer. We are using Random Forest, Decision Tree, Logistic Regression, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) models for detection. And using the Wisconsin breast cancer diagnostic data set for predictive analysis. We will use 70% of the data set for learning the model and remaining 30 % dataset will be used for testing the model.

The attributes we using are ID number, Diagnosis, ten real-valued features are computed for each cell nucleus (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension). By using this we will calculate the accuracy of each model for both train data set and test data set. Then we will compare the accuracy of each model.

After training the Random Forest, Decision Tree, Logistic Regression, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) models with train data set we got accuracies 94.724%, 96.985%, 88.693%, 87.688%, 87.437% and with test data set we got accuracies 95.906%, 97.076%, 90.058%, 87.719%, 84.211% respectively.

So, after analysing the five models we came to conclude that the Decision Tree mode is the best we got 97% accuracy and second-best model is Random Forest we got 95% accuracy.

**INTRODUCTION**

Breast Cancer is one of the major health issues occurring in women. It is caner which starts in the breast tissue. Major problem with this disease is that we are still unknown with the actual cause of this cancer and scientist are working on it day n night.

Some of the commonly known symptoms that are being used these days to detect the breast cancer are-

1. Breast rash
2. Change is the size, shape, or appearance of the breast
3. Occasional Discharge from the nipple
4. A lump or mass in the breast that feels uneven from the surrounding tissue
5. Changes in the skin over the breast
6. Mild or sever Breast pain
7. Inverted or pulling-in of the nipple
8. Abrupt scaling and peeling of skin around the darker region of nipple.

We as a Machine Learning and Artificial Intelligence enthusiastic have come up with some model comparisons that can be used to detect the breast cancer based on the breast scan that contains some of the features such as cell nucleus (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension) which we are using to train our model. These are based on the common symptoms that are being used by the doctors to predict the breast cancer.

We are using this model so that we can come up with a best model with high accuracy so that it can be used in real life medical field.

**METHODOLOGIES AND ALGORITHM**

A **decision tree** is a classification technique that divides a population into segment of branches to estimate the decision that could be taken from a particular path being followed building an inverted tree consisting of a root node, interior nodes, and leaf nodes. If your sample size is very large, you can split it into a training dataset and a validation dataset. The training data set is used to build a decision tree model and the validation data set is used to determine the appropriate tree size needed to archive the optimal final model.

**Random Forest** is an technique that is mainly used for classification and regression. But we only use classification. Random forest consists of multiple decision tree. In random forest there we divide the datasets into chunks and pass each chunk to different decision tree and the we classify the output data according to binary value.



The following steps explain how the random forest algorithm works.  
Step 1: Select a random sample from the given set of data.  
Step 2: Decision tree will be build for all the datasets.  
Step 3: Voting is done by averaging the decision tree.  
Step 4: Selection is done on the basis of maximum vote achieved.

**Logistic Regression** is a machine learning algorithm based upon Probability used for classification problems, a predictive analysis algorithm. We can call logistic regression a linear regression model, but logistic regression uses a more complex cost function whereas it is not used in linear regression. This cost function is also called "sigmoid function" or "logistic function" instead of linear function.  
  
The logistic regression hypothesis tends to constrain the cost function between 0 and 1. So no linear function can represent it. This is because it can have values ​​greater than 1 or less than 0, which are impossible according to the logistic regression hypothesis.

***Sigmoid function:*** f(x) =

***Hypothesis representation:***

***Cost function:***

***Gradient decent function:***



A **Support Vector Machine** (SVM) is a supervised machine learning algorithm which is not only used for regression but also for classification. We're talking about regression problems too, but it's best suited for classification.

Basically, SVM with the help of its hyperplane features able to create a separation plane between the characteristics dividing them into classes. These classes use a boundary decision plane named hyperplane.



**Multilayer Perceptron (MLP)** is fully connected extension of feedforward neural networks. It consists of three types of layers: input layer, hidden layer and output layer.

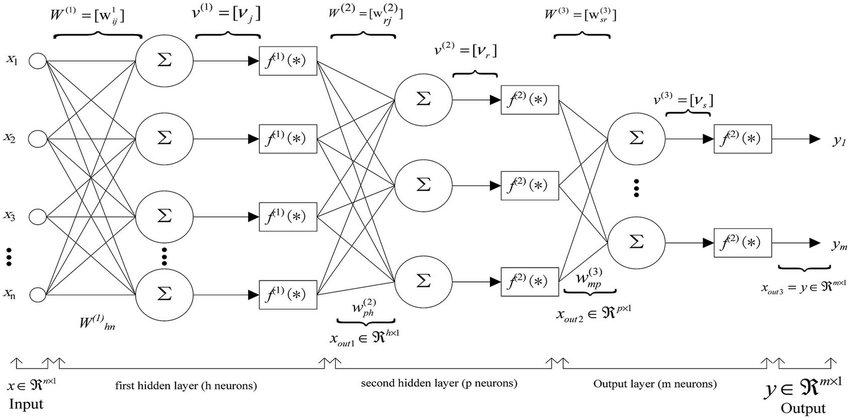
The input layer receives the input from the previous neuron or any other sources.

Any number of hidden layers sandwiched between the input and output layers is the actual working engine of MLP used in all computational task.

Task such as error examination , classification and prediction is performed at output layer. Data flows forward from the input layer to the output layer, similar to the feedforward network in MLP. MLP neurons are trained using a backpropagation learning algorithm in which the error obtained are propagated back to each node to update the weigh before the next epoch begin. MLPs are designed to approximate arbitrary continuous functions and solve problems that are not linearly separable. The primary use cases for MLP are pattern classification, recognition, prediction, and approximation.

The computation taking place at every neuron in the output and hidden layer are as follows,

Includes displacement vectors b (1), b (2); Weight matrices W (1), W (2) and activation functions G and s. The set of parameters to study is the set θ = {W (1), b (1), W (2), b (2)}. Common choices for s include the tanh function with  or the logistic with  . The sigmoid function is included.



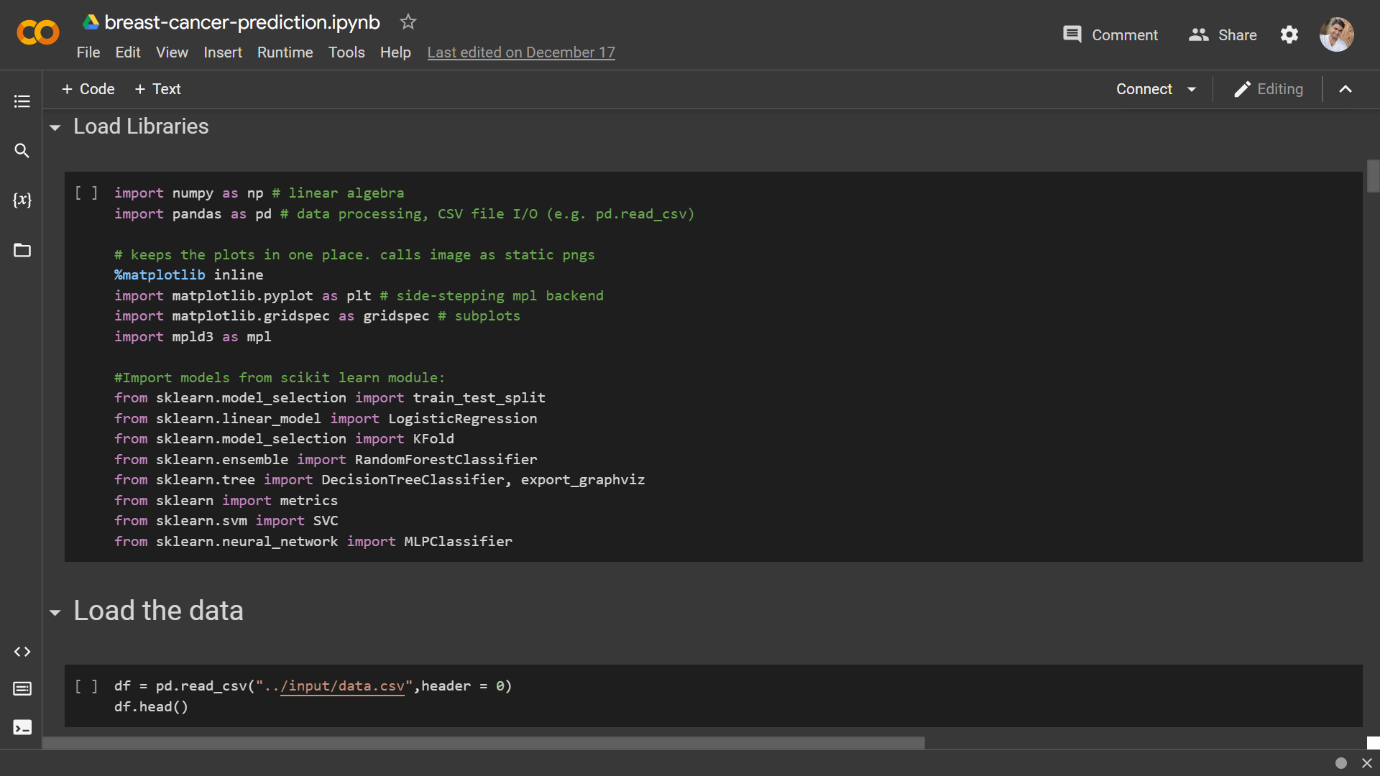
**HARDWARE REQUIREMENTS**

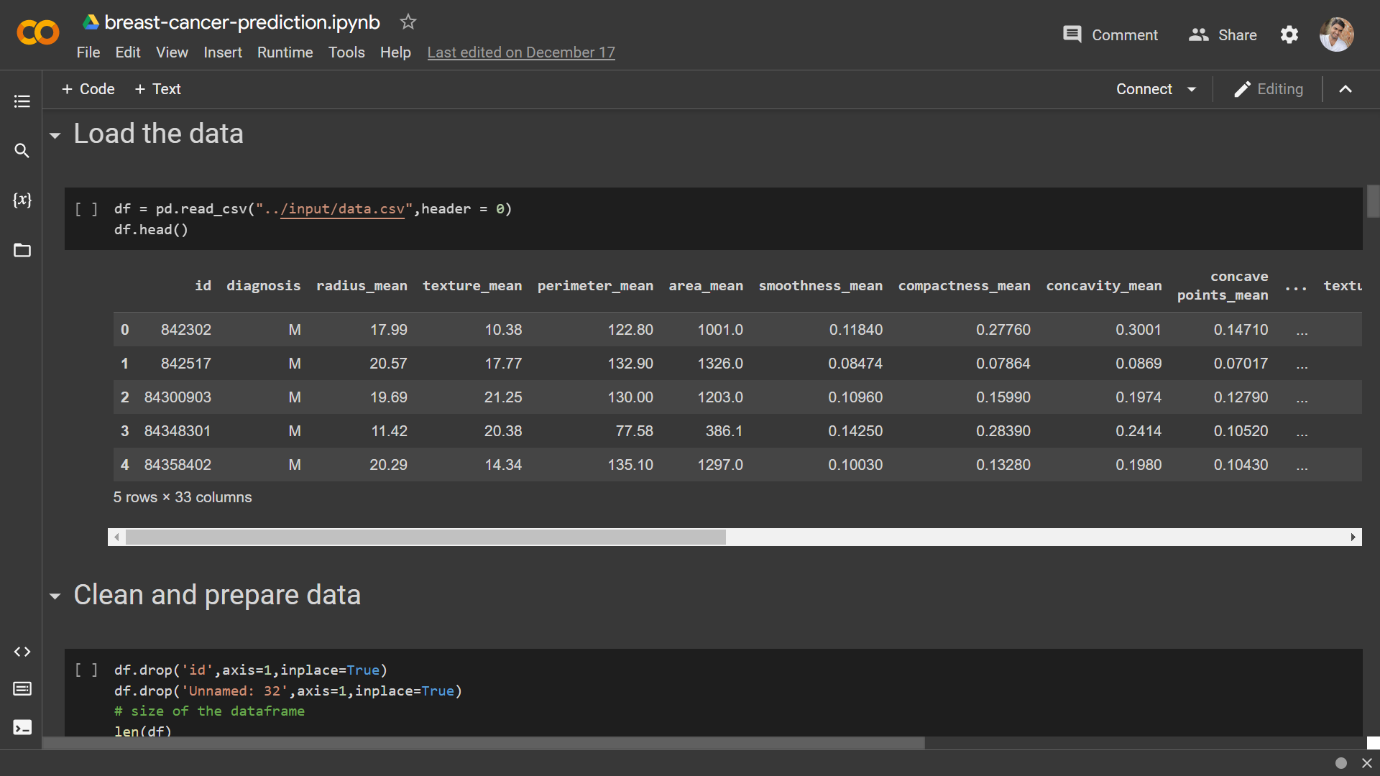
* Processor : Any Processor above 500 Megahertz
* RAM : 4 GB
* Hard Disk : 500 GB
* System : Intel i3 2.4 GHz

**SOFTWARE REQUIREMENTS**

* Operating System : Windows 10 or Above
* Programming language : Python
* IDE\* : Jupyter Notebook
* Tools\* : Anaconda
* \*Preferred Software : Google Collab or Kaggle

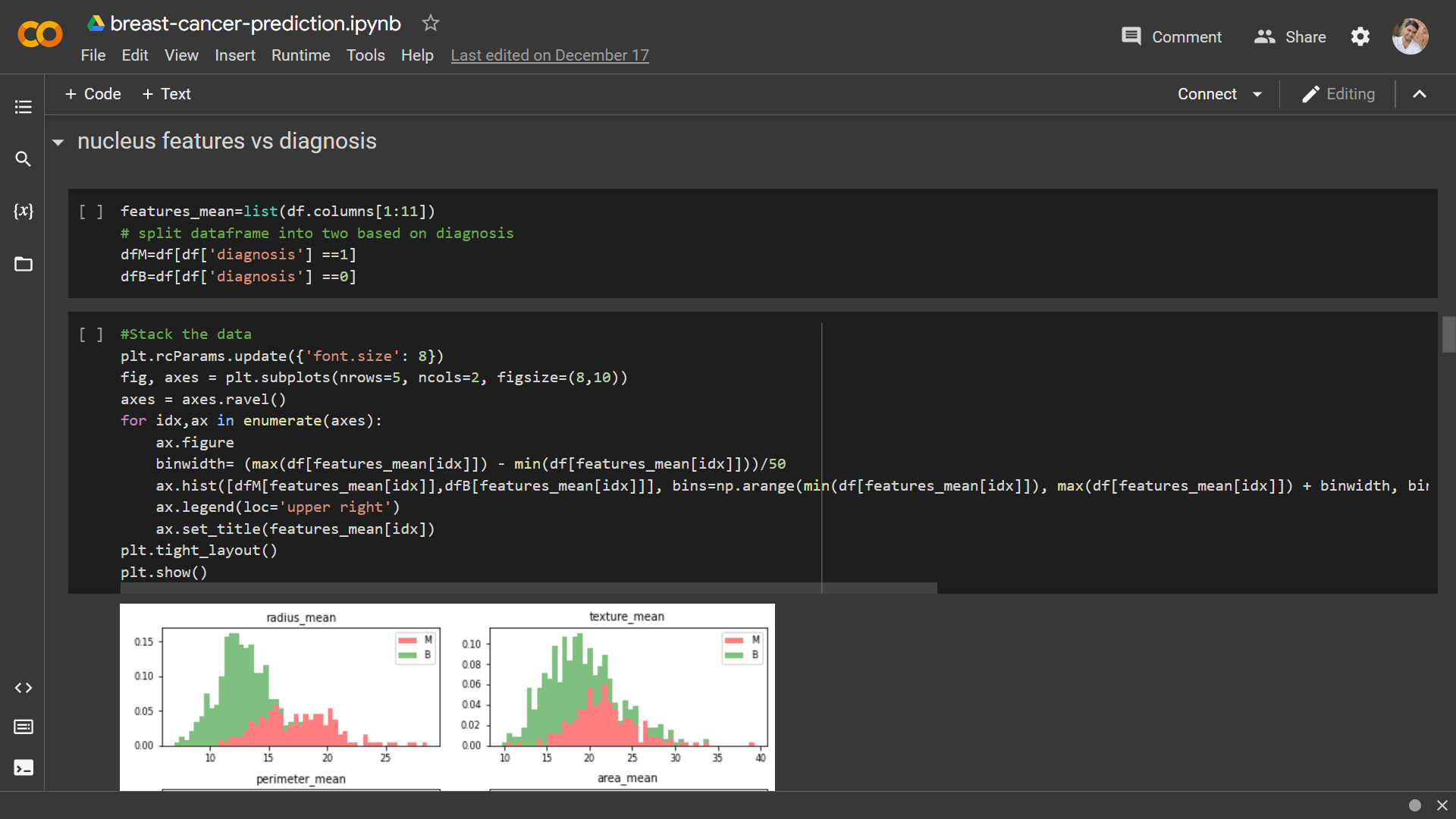
**RESULT AND DISCUSSION**

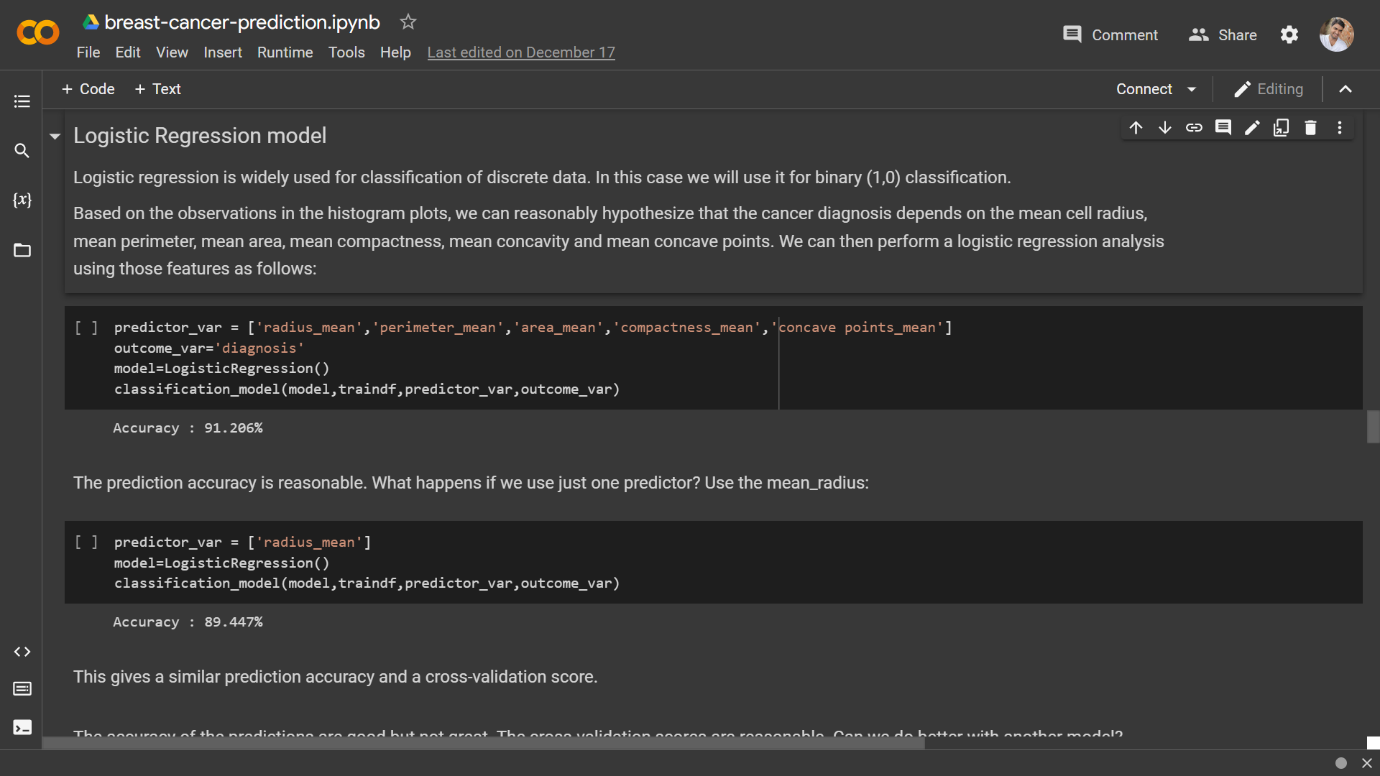
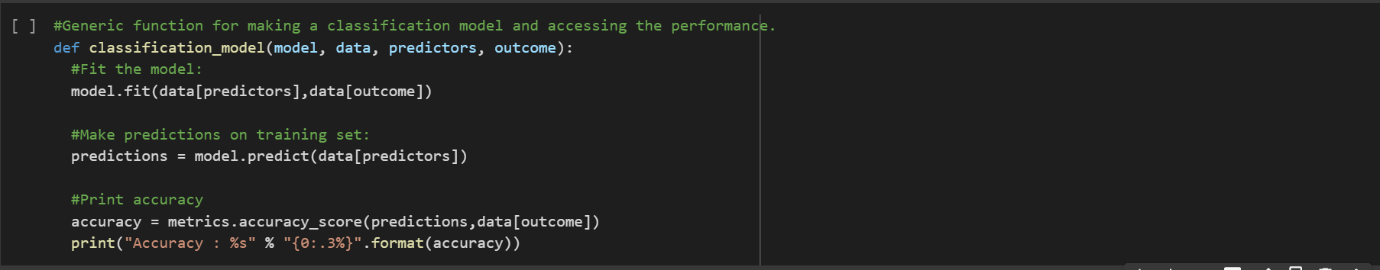


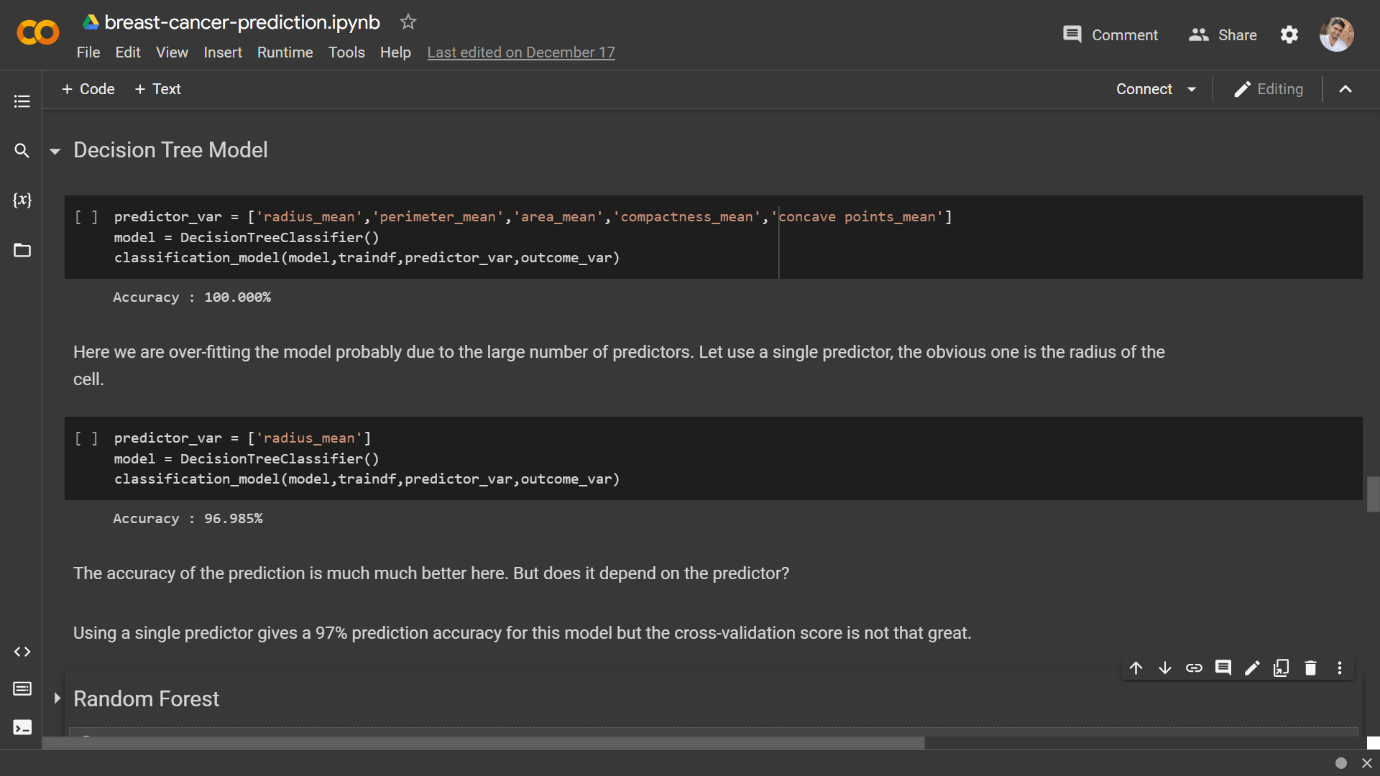


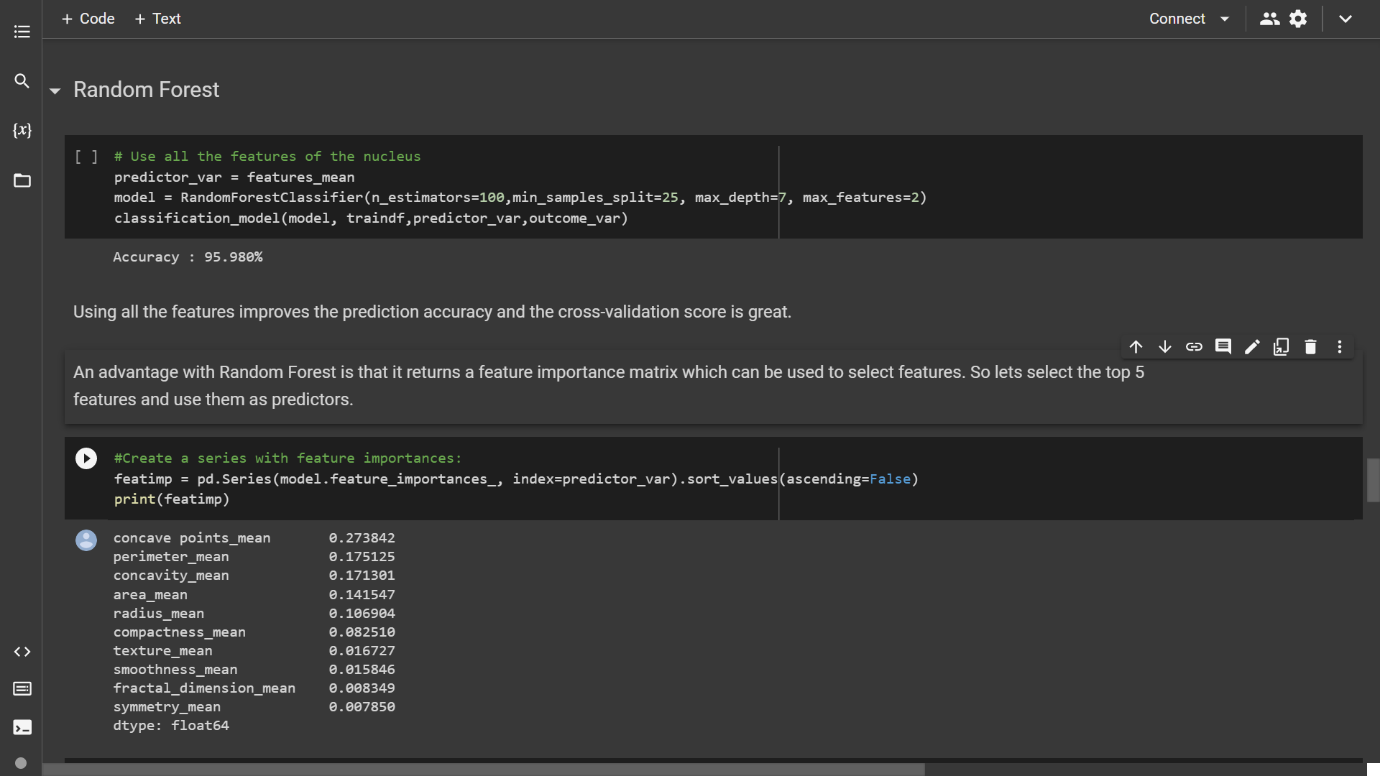


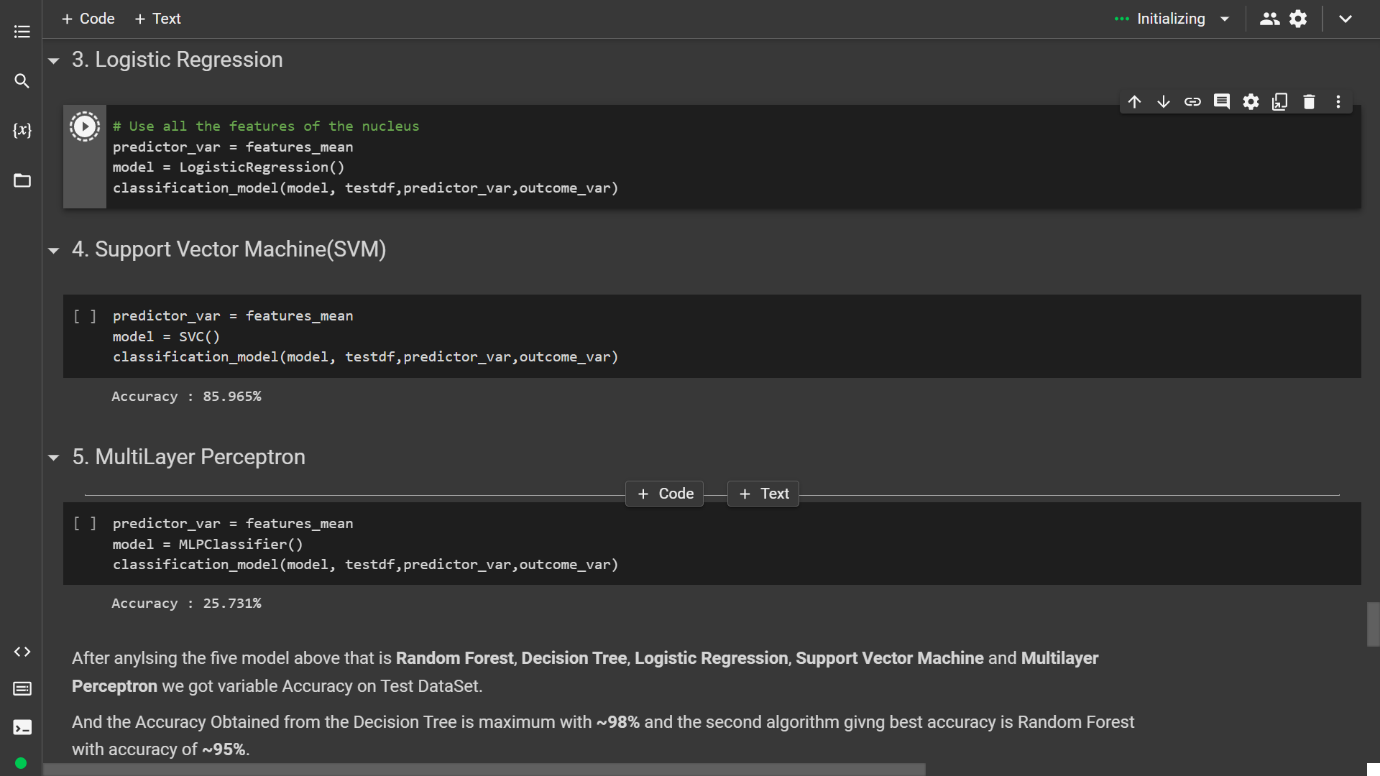
Observations

1. Mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in classification of the cancer. Larger values of these parameters tends to show a correlation with malignant tumors.
2. Mean values of texture, smoothness, symmetry or fractual dimension does not show a particular preference of one diagnosis over the other. In any of the histograms there are no noticeable large outliers that warrants further cleanup.

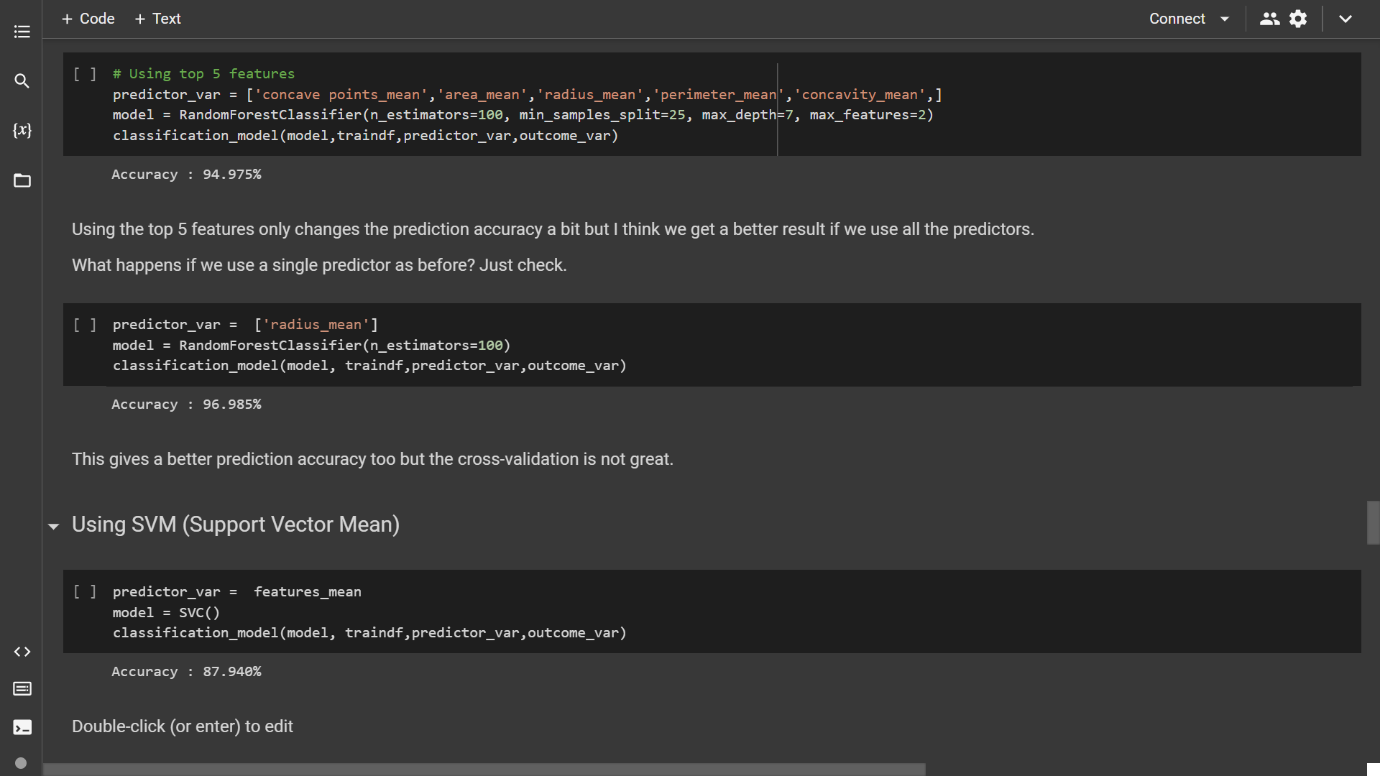
****







**CONCLUSION**

After analysing the five model above that is Random Forest, Decision Tree, Logistic Regression, Support Vector Machine and Multilayer Perceptron we got variable Accuracy on Test DataSet.

And the Accuracy Obtained from the Decision Tree is maximum with ~98% and the second algorithm giving best accuracy is Random Forest with accuracy of ~95%.

Now on the basis of the Accuracy obtained I have plotted the graph.

**FUTURE WORK**

The current model is able to show us which model we can use to detect the presence of Breast Caner by analysing the details of the breast tissue available. Our scope that was build while making this project is to further increase our accuracy from 95 % to 98-99% so that is could detect the patient more precisely. Along with that we would in future will train our model on different algorithms and will try to find best algorithm.

**REFERNCES**

1. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>
2. <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>
3. [Dr. Geeta Kadayaprath - Cancer Care / Oncology, Book Online Appointment, Video Consultation, Check OPD Timings, View Fees | Max Hospital (maxhealthcare.in)](https://www.maxhealthcare.in/doctor/dr-geeta-kadayaprath)
4. <https://www.cdc.gov/cancer/breast/basic_info/symptoms.htm>

**THANK YOU !!!**