

BREAST CANCER PREDICTION

## CS 675 Machine Learning Milestone 4

Abstract

|  |  |  |  |
| --- | --- | --- | --- |
| The report is on | using multiple different models to | | identify which |
| model is | more accurate | for diagnosing | breast cancer. |
| Dhruvi Makadia, | Maryana | Stefanyshyn, & | Ashween |
|  |  |  | Sudan |

ddm8@njit.edu, ms458@njit.edu, & aks69@njit.edu

**Table of Contents**

[**BACKGROUND OF DATASET** 2](#_Toc11285)

[**OVERVIEW OF DATASET FROM MILESTONE 2** 2](#_Toc11286)

[Linear regression model 3](#_Toc11287)

[DECISION TREE MODEL 3](#_Toc11288)

[RANDOM FOREST 4](#_Toc11289)

[**MILESTONE 4** 4](#_Toc11290)

[K NEIGHBORS CLASSIFIER (KNN) 4](#_Toc11291)

[SUPPORT VECTOR MACHINE (SVM) 9](#_Toc11292)

[PRINCIPAL COMPONENT ANALYSIS (PCA) 12](#_Toc11293)

[**CONCLUSION** 14](#_Toc11294)

[**TASK BREAKDOWN ASSIGNMENT** 14](#_Toc11295)

[**REFERENCES** 14](#_Toc11296)

# Background of Dataset

Breast cancer is the most common malignancy among women and is the second leading cause of cancer death among women. Breast Cancer occurs as a result of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. A tumor does not mean cancer - tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous).

This dataset is a collection of testing data from the breast fine needle aspiration (FNA) test. It is a quick procedure to perform, which removes some of the fluids or cells from a breast lesion or cyst with a fine needle.

Expected Outcome:

The goal is to classify a breast cancer tumor using two training classifications and predict if it is malignant or benign.

* 1 = Malignant (Cancer is present)
* 0 = Benign (Cancer is absent)

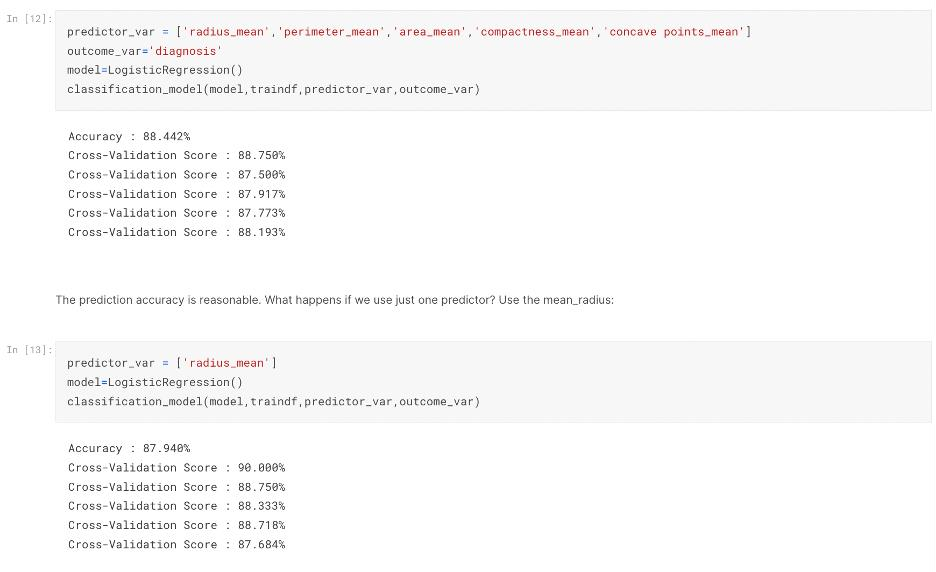
# Overview of Dataset from Milestone 2

The report Breast Cancer Prediction dataset by Buddhini Waidyawansa. Buddhini used the Breast Cancer Wisconsin (Diagnostic) data set which is the same dataset used throughout the report to identify the top best model with the highest accuracy that should be used for diagnosing breast cancer. The peer review from milestone 2 used the Breast Cancer Prediction dataset by Buddhini Waidyawansa. Buddhini did exploratory data analysis on the dataset using multiple predictive and classification algorithms like Logistic Regression, Decision Tree, and Random Forest. Her analysis demonstrated the mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. Also, Buddini showed the difference between nucleus features vs diagnosis. Presenting the graphs for perimeter, texture mean, concave mean, fractional dimension mean etc . Using Linear regression, she hypothesized that the cancer diagnosis depends on the mean cell radius, mean perimeter, mean area, mean compactness, mean concavity, and mean concave points. Then, she performed a logistic regression analysis using those features.

# Linear regression model

Buddhini hypothesized that the cancer diagnosis depends on the mean cell radius, perimeter, area, compactness, concavity, and concave points. As presented below in the screenshot logistic regression analysis has been performed: Using the linear regression model accuracy came to

88.442%.



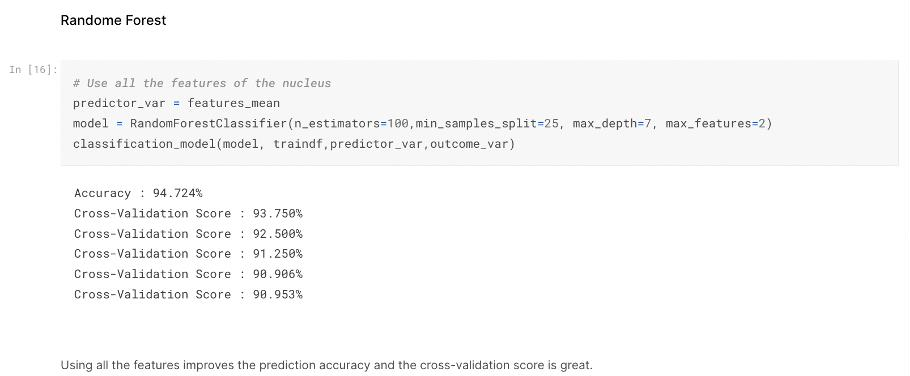
# Decision Tree Model

Buddhini hypothesized that the cancer diagnosis depends on the mean cell radius, perimeter, area, compactness, and concave points. As presented below in the screenshot decision tree model analysis has been performed: Using the decision tree model accuracy came to 100%.



# Random Forest

Buddhini hypothesized that the cancer diagnosis depends on the mean cell radius, perimeter, area, compactness, and concave points. As presented below in the screenshot of random forest model analysis has been performed: Using the random forest model accuracy came to 94.724%.



According to Buddhini, the top three models used in her Breast Cancer prediction analysis is a random forest with the top five predictors, “concave, points\_mean, area\_mean, radius\_mean, perimeter\_mean, and concavity\_mean”. Random forest is the model which is providing a more accurate accuracy of 94.724% than the decision tree or linear regression model.

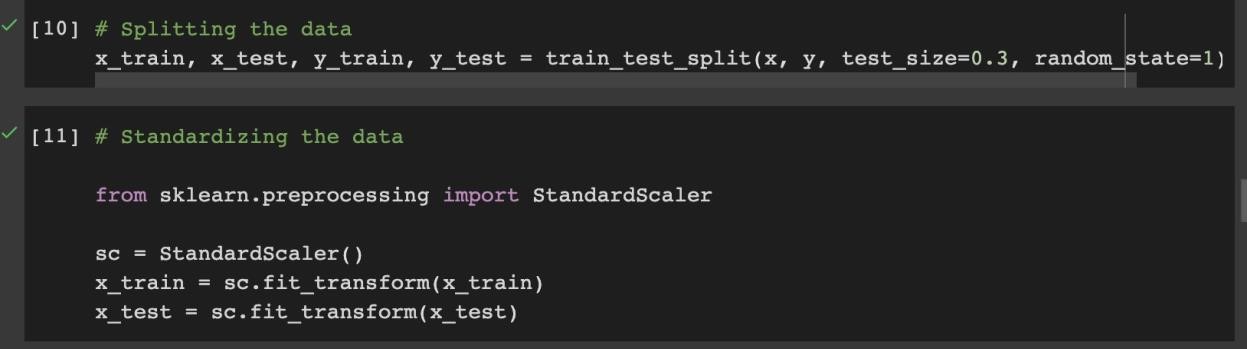
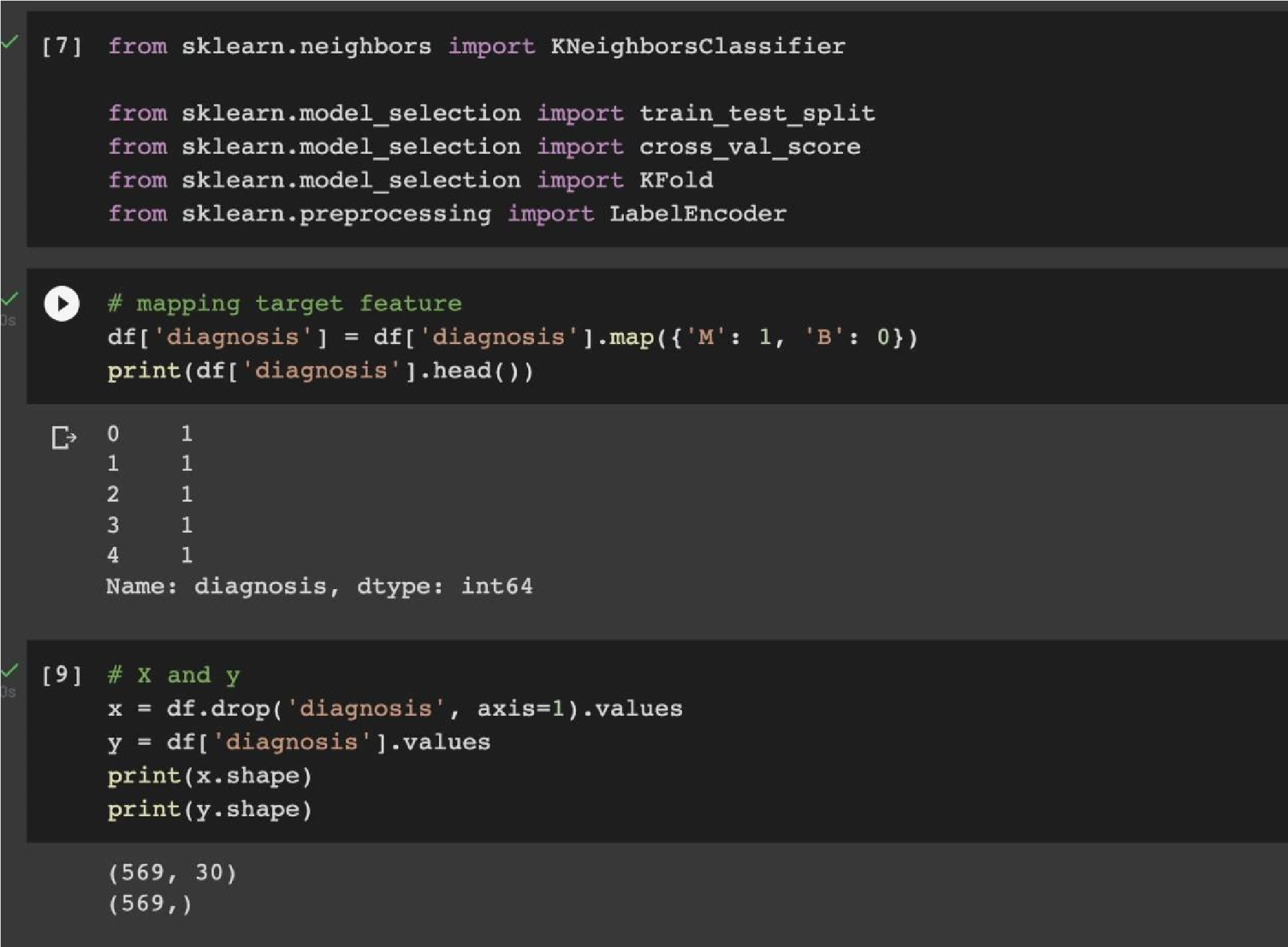
# Milestone 4

The team took the initiative and took Budhhini’s analysis to next level by adding K Neighbors Classifier, PCA, and SVM analysis to this report to identify the best model to be used for diagnosing breast cancer.

# K Neighbors Classifier (KNN)

The KNN Algorithm is an unsupervised learning algorithm that groups objects into clusters based on similarities. For the breast cancer dataset, there are 2 possible predicted classes: **"1" = malignant** (indicates presence of cancer cells) and **"0" = benign** (indicates absence of cancer cells)

First, the data is split into testing and training data and scaled it. After analyzing it, we found that there are two unique values (malignant and benign) from the “diagnosis” column in the dataset.



We graphed values of k vs. Error to find k with the smallest error. From the graph, the best value to use for k is 9 or 10.

the

We ran the KNN classifier with 10

neighbors

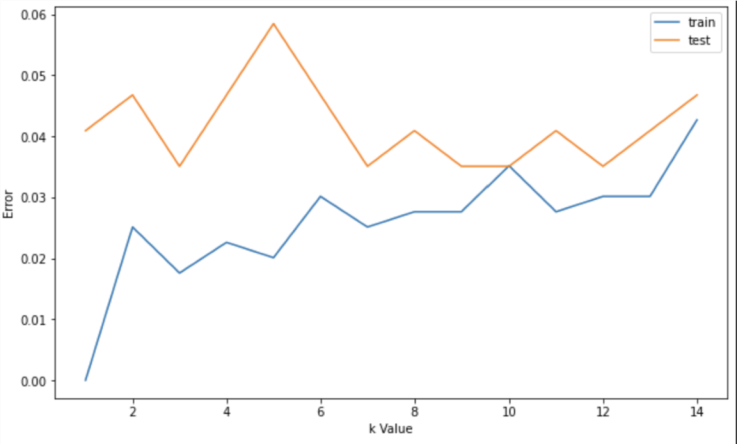
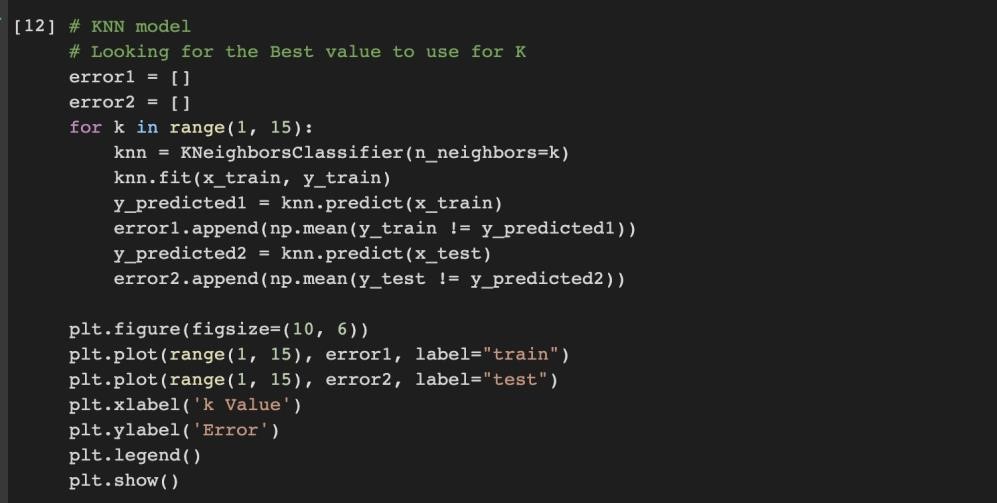
and

plotted

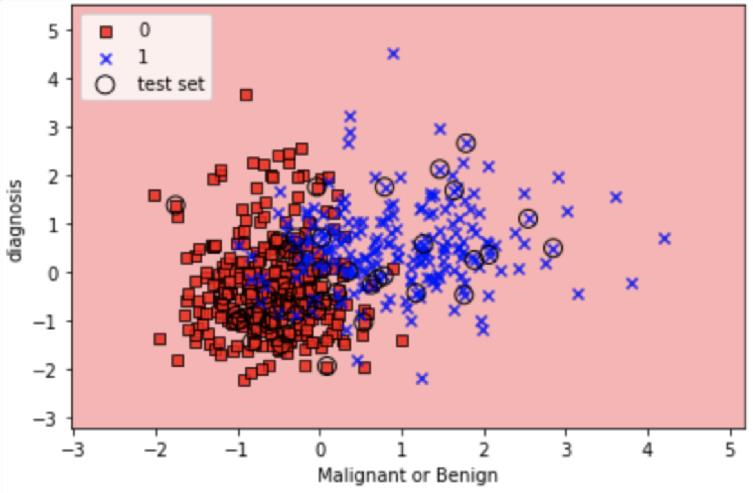
classification output. As presented

in the graph below: the color red is

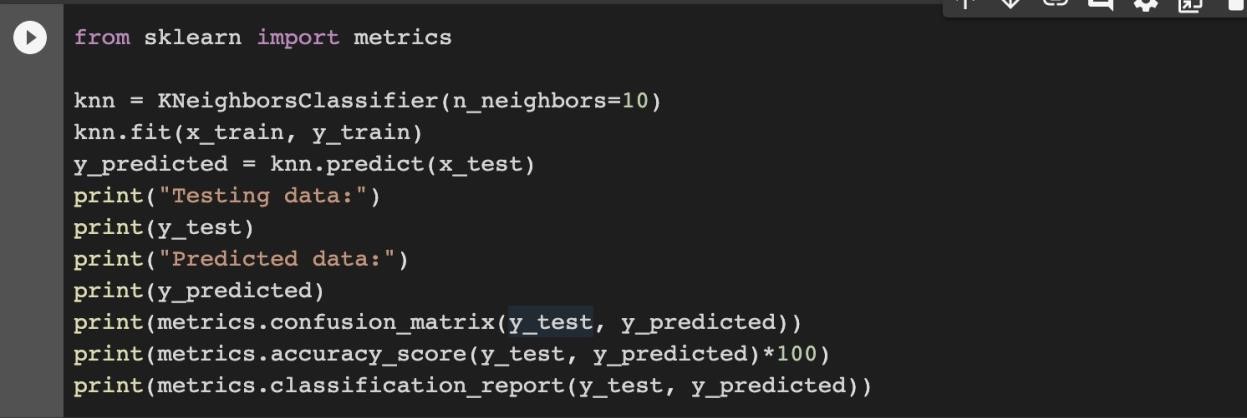
representing “benign” and blue is

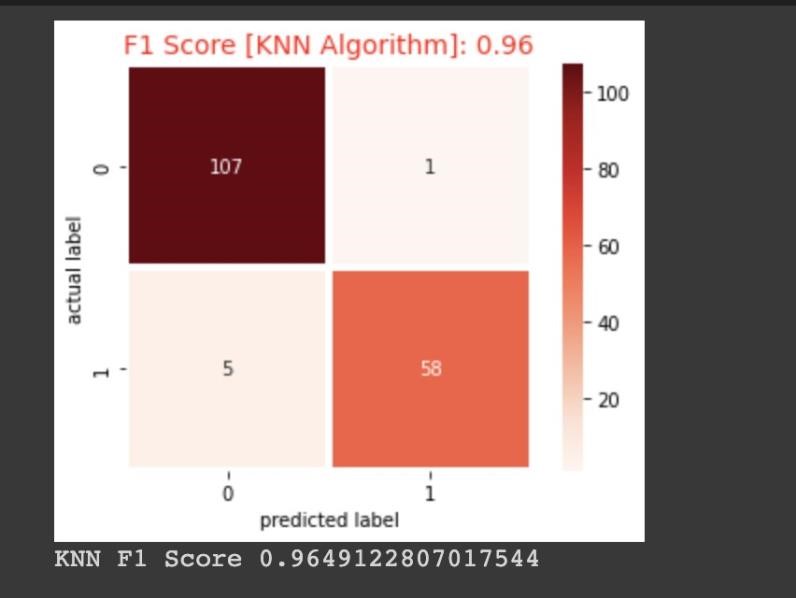
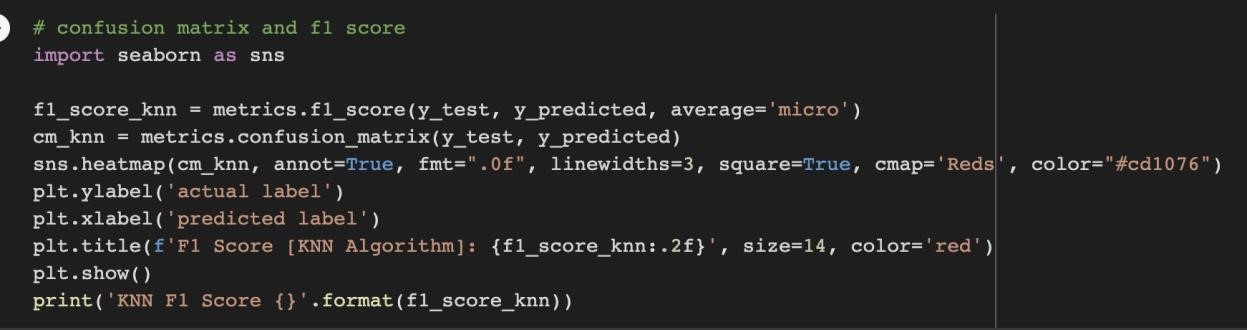
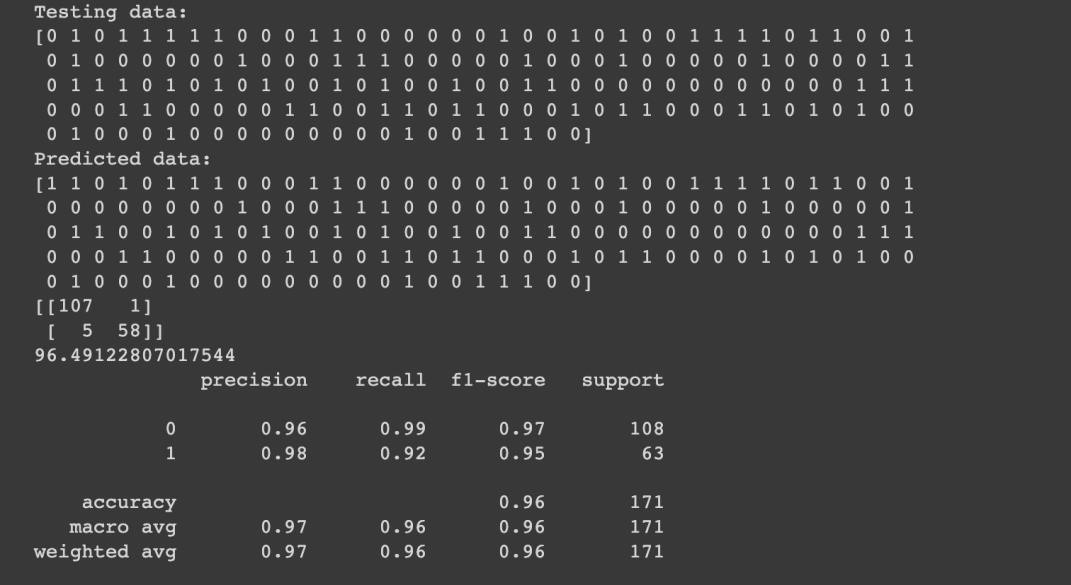


representing “malignant” cases.



Then, we fit the KNN classifier and found predicted values. The screenshot below shows the data’s actual values “Testing data” and KNN’s predicted values “Predicted data”. The accuracy score came as 96.49 %, which is very high.



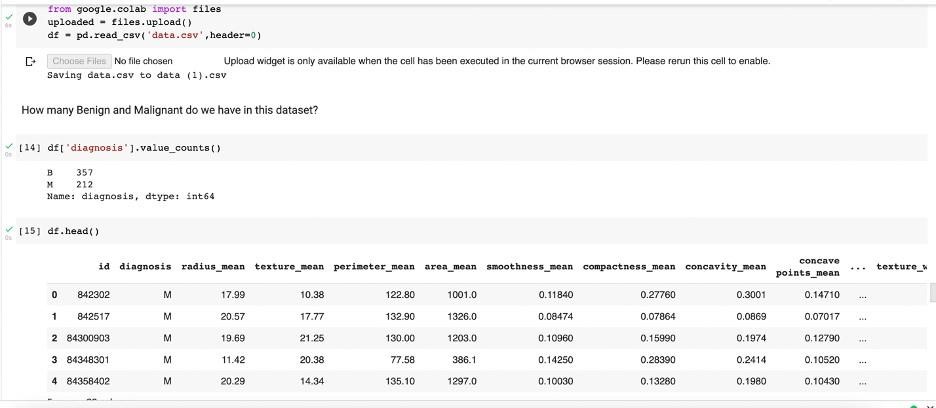


**KNN Classification Results:**

The KNN Classifier made a total of 108 predictions. A total of 171 patients were being tested for the presence of breast cancer. Out of those 171 cases, the classifier predicted “1 = malignant” 63 times and “0=benign” 108 times. Therefore, 63 patients in the sample have the disease and 108 patients do not. The F1 score of KNN is 0.96 which is very good and shows high precision and accuracy of the algorithm.

# Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning model, which uses classification algorithms for two group classification problems. SVM model goal is to have largest possible margin between decision boundary that separates the classes and the training instances. First, we created a pipeline that used LinearSVC, which will normalize the feature values using Standard Scaler.

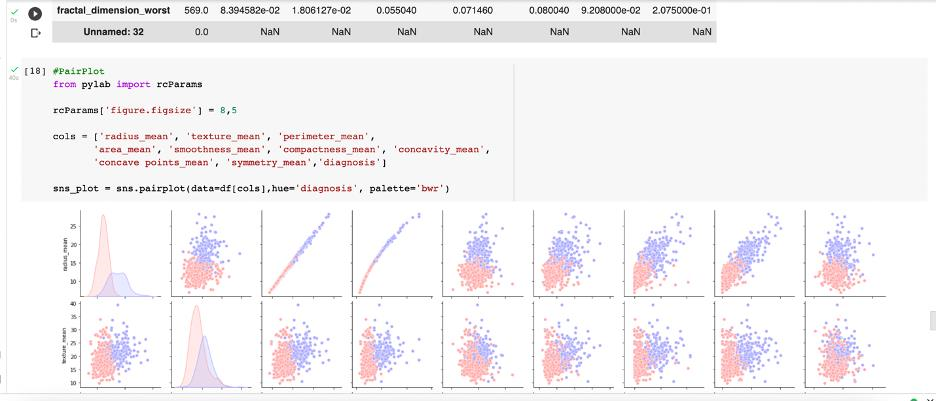


**malignant** (indicates presence of cancer cells)

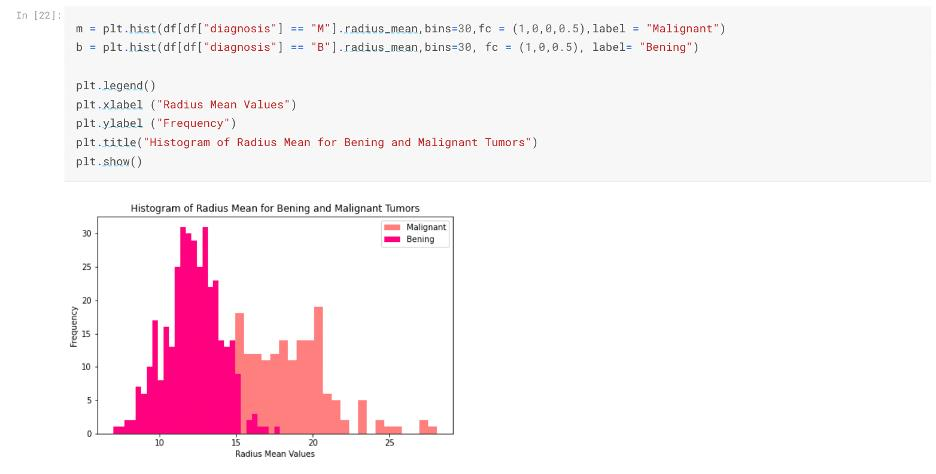
**benign** (indicates absence of cancer cells)

We performed Data Visualization with the help of Pairplot and Histogram as shown below:

From the figure below, the pairplots are shown comparing two variables between “malignant” and “benign” among radius\_mean, perimeter\_mean, area\_mean, smoothness\_mean, compactness\_mean, concavity\_mean, and concave points\_mean. This shows that “benign” is bit higher than “malignant”.



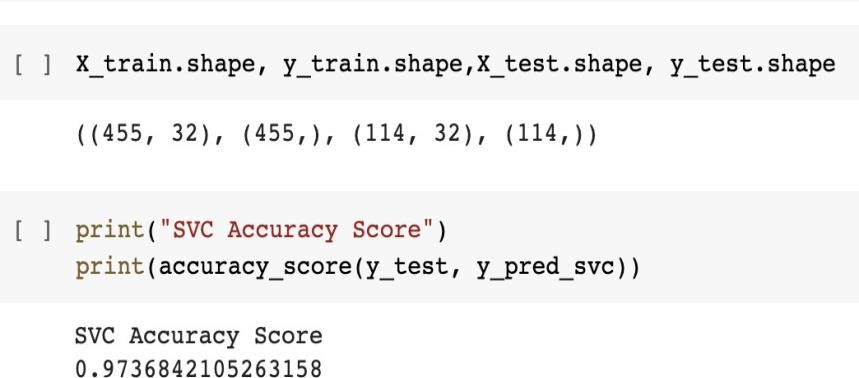
Similarly, the histogram below illustrates that there is high chance of “benign” than the “malignant”.



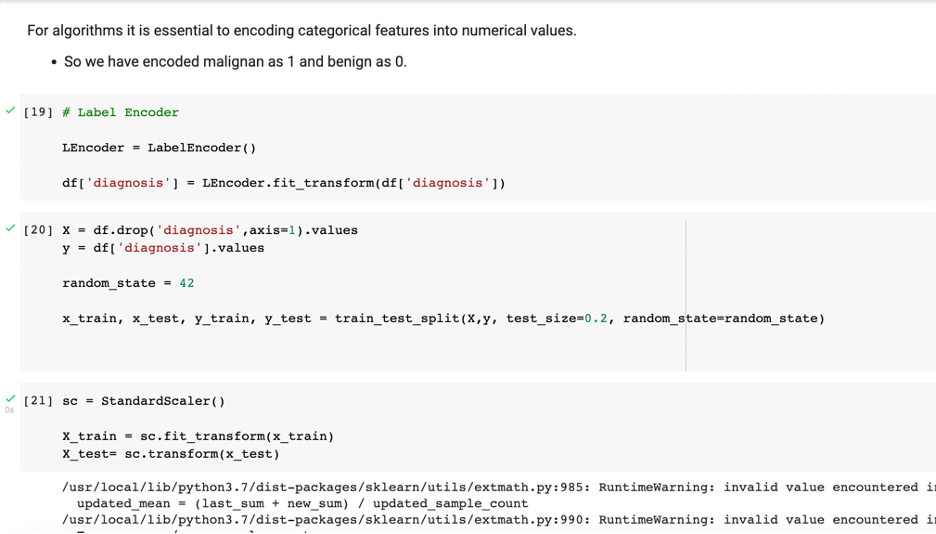
Also, the training set that we l use to train our model, it has the big part.

The testing set: is used to evaluate the performance of the model after hyperparameter tuning.

We split our data using 80% for training and the remaining 20% for testing .



SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. Then, a separator is estimated for the data. The data should be transformed in such a way that a separator could be drawn as a hyperplane.



SVM Results:

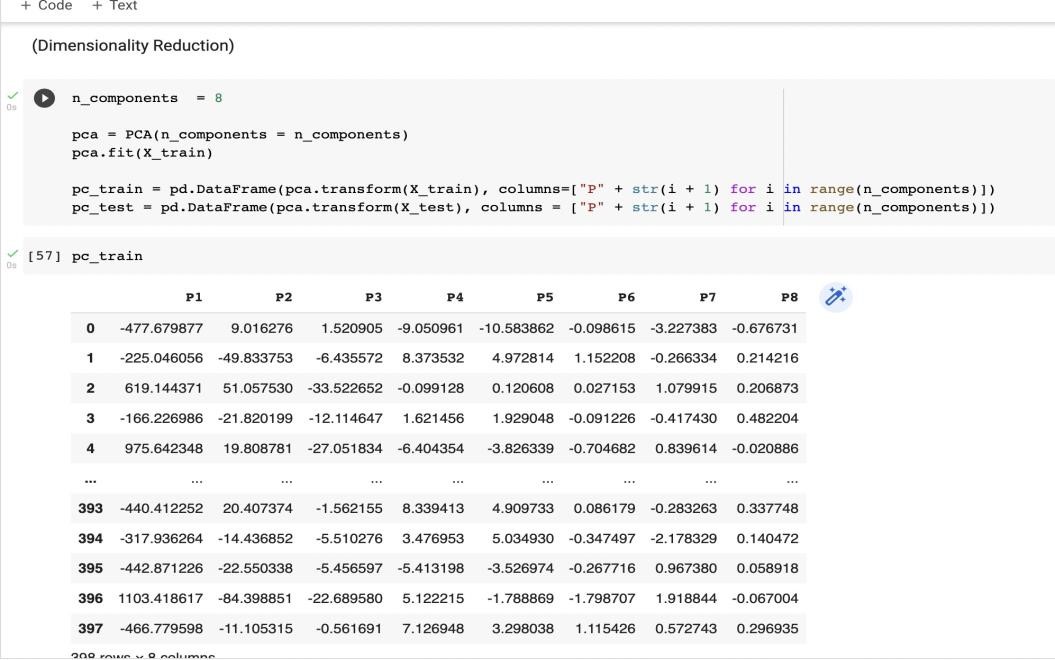


With the use of SVM algorithm, we found accuracy score of 97.368%.

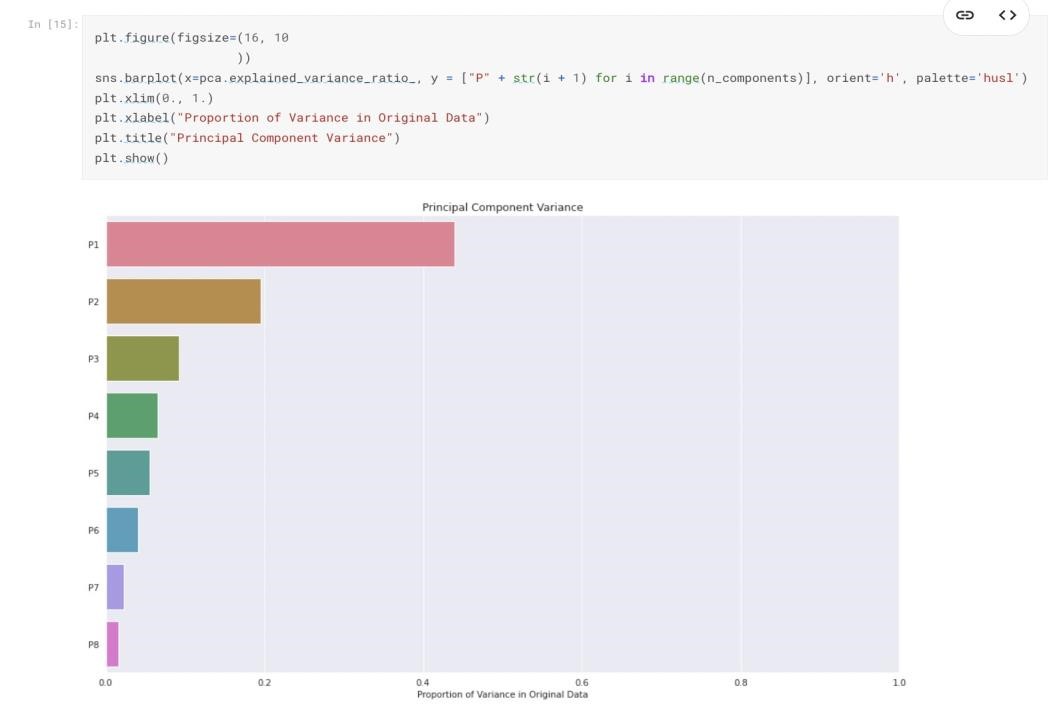
# Principal Component Analysis (PCA)

Principal component analysis (PCA) is a statistical procedure, which allows an individual to perform data analysis for predictive models. PCA is an unsupervised machine learning technique used to reduce the dimensions of the dataset. This algorithm helps in identifying relationships among different variables. PCA is a preprocessing step, which decomposes a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of the variance.

The PCA is used when we need to tackle datasets with a large number of features with different scales, some of which might be correlated. These correlations and the high dimension of the dataset bring a redundancy in the information.



Below, we graphed a plot which shows the principal component variance in proportion to variance found in original data.



PCA Results:



Applying PCA, the original features are transformed into linear combinations of new independent variables, which reduce the complexity of the dataset and thus, the computational cost.

From our results, we found the PCA algorithm to be useful and give us an accuracy score of 99.41%.

# Conclusion

**Accuracy Scores for Breast Cancer Predictions:**

Logistic Regression: 88.44%

Random Forest: 94.72%

Decision Tree: 100.00%

K Neighbors Classifiers: 96.49%

PCA: 99.41% SVM: 97.36%

Looking at the overall accuracy, the Decision Tree Model had the highest accuracy score of 100%. However, that is very rare and means there can be overfitting of the data. Our analysis found the PCA model to have the next highest accuracy score of 99.41%.

# References

**Link to Buddhini Waidyawansa’s Breast Cancer Analysis:**

https://www.kaggle.com/code/buddhiniw/breast-cancer-prediction/data.

**Link to the Breast Cancer Wisconsin (Diagnostic) Data Set:**  https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data **Link to Breast Cancer Presentation (Google Powerpoint):**

https://docs.google.com/presentation/d/1BELTxrqzIw4t4KIqzkZQvNXhnufmG3CeeJB81SKKr4/ed it#slide=id.p

**Link to our Notebook and Source Code (Google Collab Notebook):**

https://colab.research.google.com/drive/1znLtJXpPd1tgUKO4Y\_uH7qXZfyhXuHDB?authuser=2