Cancer Prediction

COURSE PROJECT REPORT

18CSE398J -Machine Learning - Core Concepts with Applications

(2018 Regulation)

III Year/ VI Semester

Academic Year: 2022 -2023 (EVEN)

By

Joel Santosh George - RA2011003010051

Kevin Thomas Koshy - RA2011003010018

Yash Mehta- RA2011030010066

Under the guidance of

Vijayalakshmi V

Professor

Department of Data Science and Business Systems

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Kancheepuram

MAY 2023

Abstract

Machine learning (ML) has become an increasingly important tool in cancer research and treatment, with the potential to improve our ability to predict cancer incidence and progression, identify high-risk individuals, and develop targeted and effective treatments.

In this report, we go through how ML algorithms can be used to process large volumes of data and identify patterns for cancer prediction that may be difficult for humans to detect.

We look at the importance of developing personalized risk scores for individuals, taking into account individual factors such as age, gender, occupational hazards etc and coming up with a possible risk level for the given parameters.

By leveraging the power of advanced algorithms and large datasets, ML is poised to help us better understand the underlying mechanisms of cancer, identify new biomarkers, and develop more personalized treatments.

Introduction

The machine learning project we have made is based on the "Breast Cancer Dataset" and it aims to develop a predictive model that can help identify individuals whose tumor parameters suggest malignant or benign cancer.. The project involves using advanced machine learning techniques such as classification algorithms to analyze the dataset's various parameters and predict the likelihood of cancer in patients. By doing so, the project aims to provide healthcare professionals with a tool that can help them diagnose cancer at an early stage and provide timely treatment, improving patient outcomes.

The project's objective is to develop a robust and accurate model that can take in the patient's input parameters and predict the likelihood that it is a malignant or benign cancer. This can help healthcare professionals make better decisions regarding diagnosis and treatment, ultimately improving patient outcomes and reducing the burden of cancer on society. The project requires extensive data preprocessing, feature selection, and model training using advanced ML algorithms, making it a challenging. Overall, this project has the potential to make a significant impact on cancer diagnosis and treatment, paving the way for more effective preventive measures and personalized medicine. Moreover the project has been deployed using Flask so as to improve ease of use and grants better functionality.

About the Dataset

The "Breast Cancer Data" is a dataset that contains various parameters related to cancer patients such as radius_mean, texture_mean, area_mean, compactness, concavity, fractal_dimensions and various other sub parameters with their mean and worst values

In breast cancer data, "worst" and "mean" can refer to a variety of features that are used to assess the severity and aggressiveness of the cancer.

This dataset can be utilized to develop machine learning models to predict cancer based on the input parameters. By analyzing the various parameters in the dataset, ML models can be trained to predict the likelihood of malignant or benign breast cancer in patients, which can be used to identify individuals who are at higher risk of developing the disease and provide them with timely medical intervention. The dataset can be an essential tool for researchers and healthcare professionals to gain insights into cancer development and design more effective preventive and treatment measures.

Method To Create the Model

- Data Cleaning and Preprocessing: The first step in the implementation process would be to clean and preprocess the dataset. This involves removing any missing or duplicate values, converting categorical variables into numerical values, and scaling the data.
- Feature Selection: The next step would be to select the most relevant features that can help in predicting the type of tumor. This can be achieved using statistical methods such as correlation analysis or machine learning-based feature selection algorithms.
- Model Selection: Once the features have been selected, the next step would be to choose an appropriate machine learning model that can predict cancer likelihood accurately. This can be done by comparing the performance of various models such as decision trees, logistic regression, support vector machines, and neural networks. For this model, we found Logistic Regression gave us the best accuracy.
- Model Training: After selecting the model, the next step would be to train the model on the dataset. This involves dividing the data into training and testing sets, tuning the model's hyperparameters, and optimizing the model's performance.
- Model Evaluation: Once the model has been trained, the next step would be to evaluate its performance using various metrics such as accuracy, precision, recall, and F1-score. This can help determine if the model is performing well or needs further improvement.
- Deployment: After evaluating the model's performance, the final step would be to deploy the model in a real-world setting. This can involve integrating the model with an existing healthcare system or developing a new software application for cancer diagnosis.

- Continuous Improvement: Finally, the implementation process should include a continuous improvement phase, where the model's performance is monitored, and the model is updated periodically to incorporate new data and improve its accuracy.

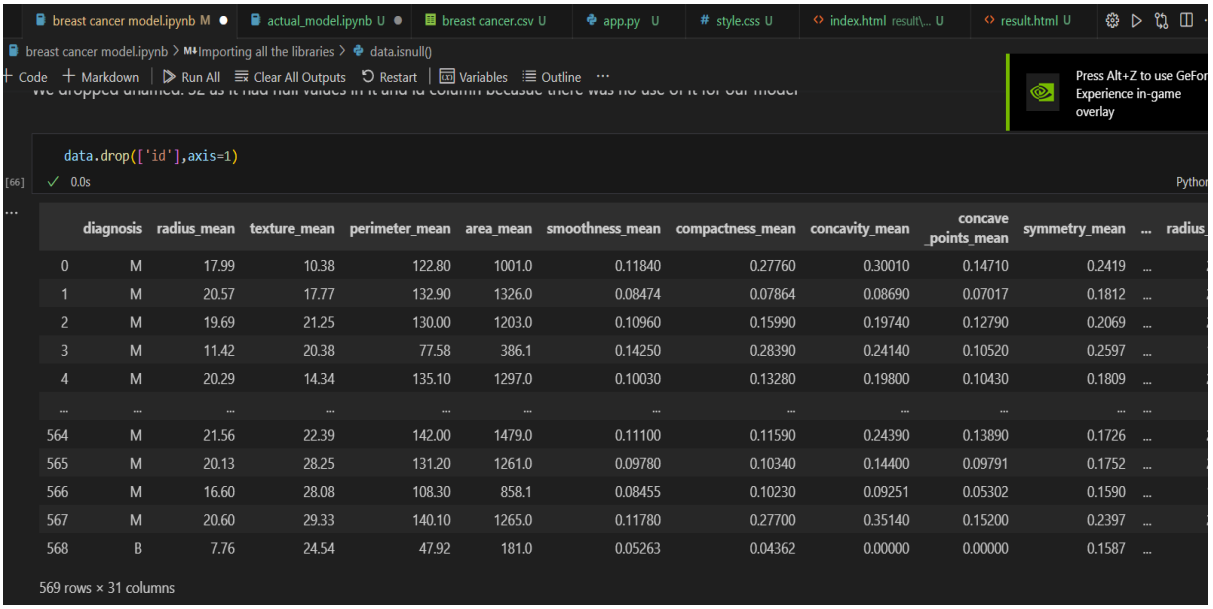
About the Algorithm

Since we are dealing with a binary classification problem, and after comparing the results with other algorithms, we decided to go with the Logistic Regression algorithm.

Logistic regression is a statistical method used to analyze the relationship between a categorical dependent variable and one or more independent variables. It is a type of regression analysis where the outcome variable is binary or dichotomous, meaning it can only take one of two possible values.

The logistic regression model estimates the probability of the dependent variable (which is typically a categorical variable) being in one of the two possible outcomes based on the values of the independent variables. It uses a logistic function (sigmoid function) to map the input variables to the output probability and determine the result.

Code



breast cancer model.ipynb M | actual_model.ipynb U | breast cancer.csv U | app.py U | style.css U | index.html result\... U | result.html U

breast cancer model.ipynb > Importing all the libraries > data.isnull()

+ Code | + Markdown | Run All | Clear All Outputs | Restart | Variables | Outline

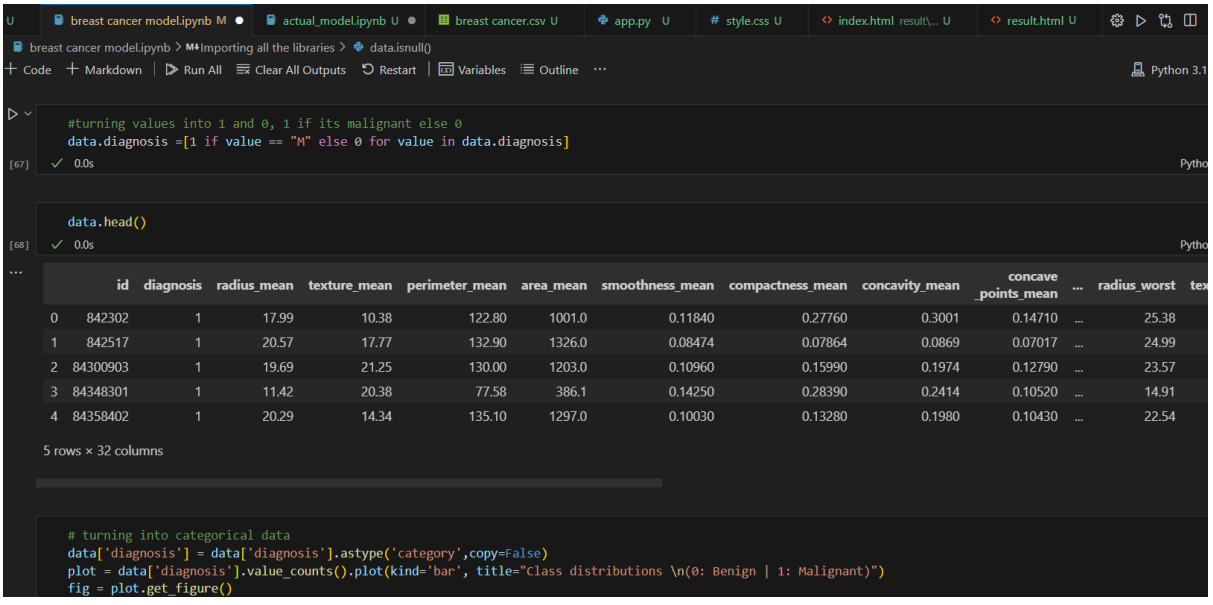
Press Alt+Z to use GeForce Experience in-game overlay

```
data.drop(['id'],axis=1)
```

[66] ✓ 0.0s

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave _points_mean	symmetry_mean	...	radius_
0	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	...	
1	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	...	
2	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	...	
3	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	...	
4	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	...	
...	...	...	...	...	...	...	...	...	...	...	...	
564	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	...	
565	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	...	
566	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	...	
567	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	...	
568	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	...	

569 rows × 31 columns



breast cancer model.ipynb M | actual_model.ipynb U | breast cancer.csv U | app.py U | style.css U | index.html result\... U | result.html U

breast cancer model.ipynb > Importing all the libraries > data.isnull()

+ Code | + Markdown | Run All | Clear All Outputs | Restart | Variables | Outline

Python 3.1

```
#turning values into 1 and 0, 1 if its malignant else 0
data.diagnosis =[1 if value == "M" else 0 for value in data.diagnosis]
```

[67] ✓ 0.0s

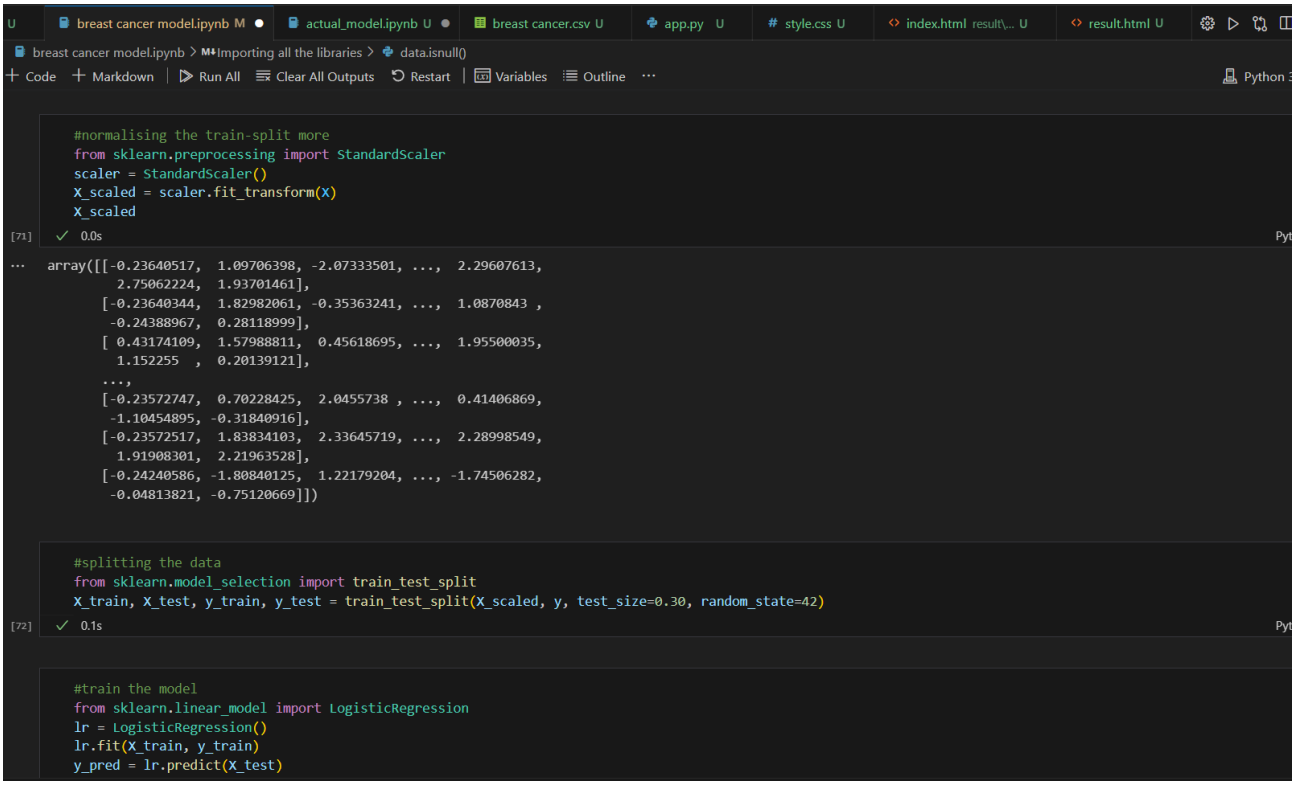
```
data.head()
```

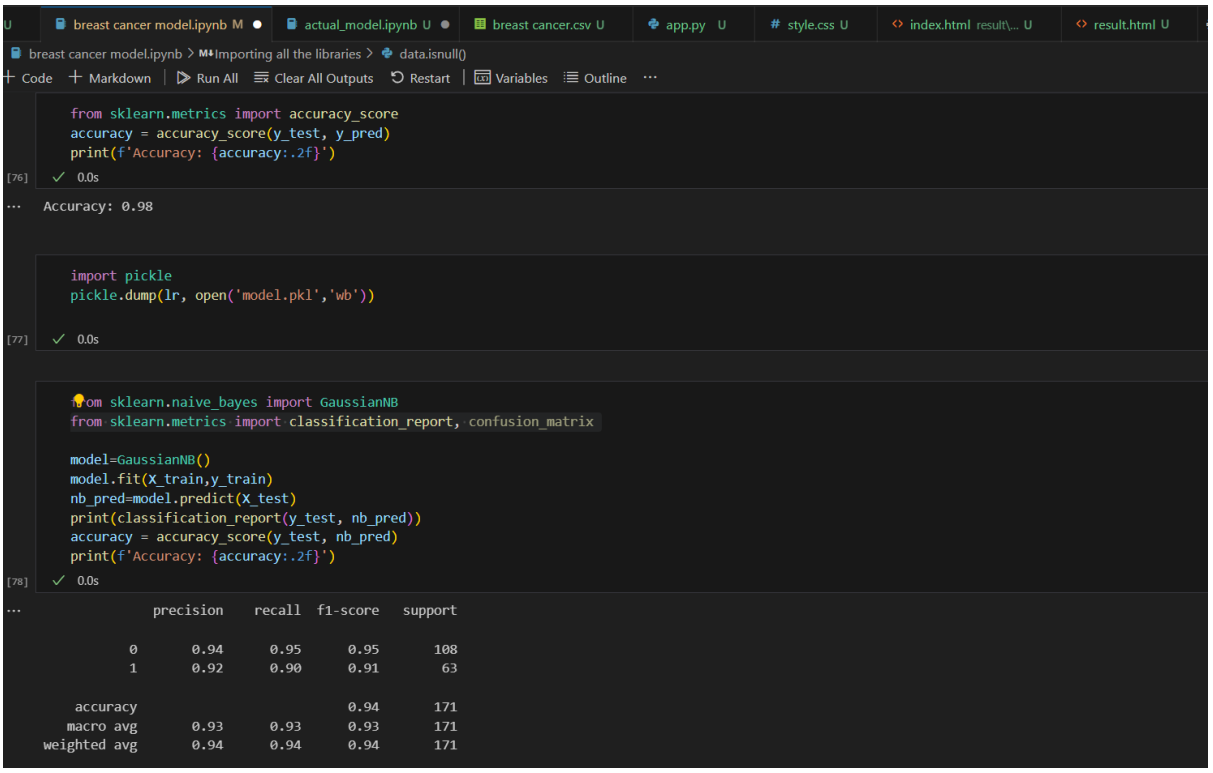
[68] ✓ 0.0s

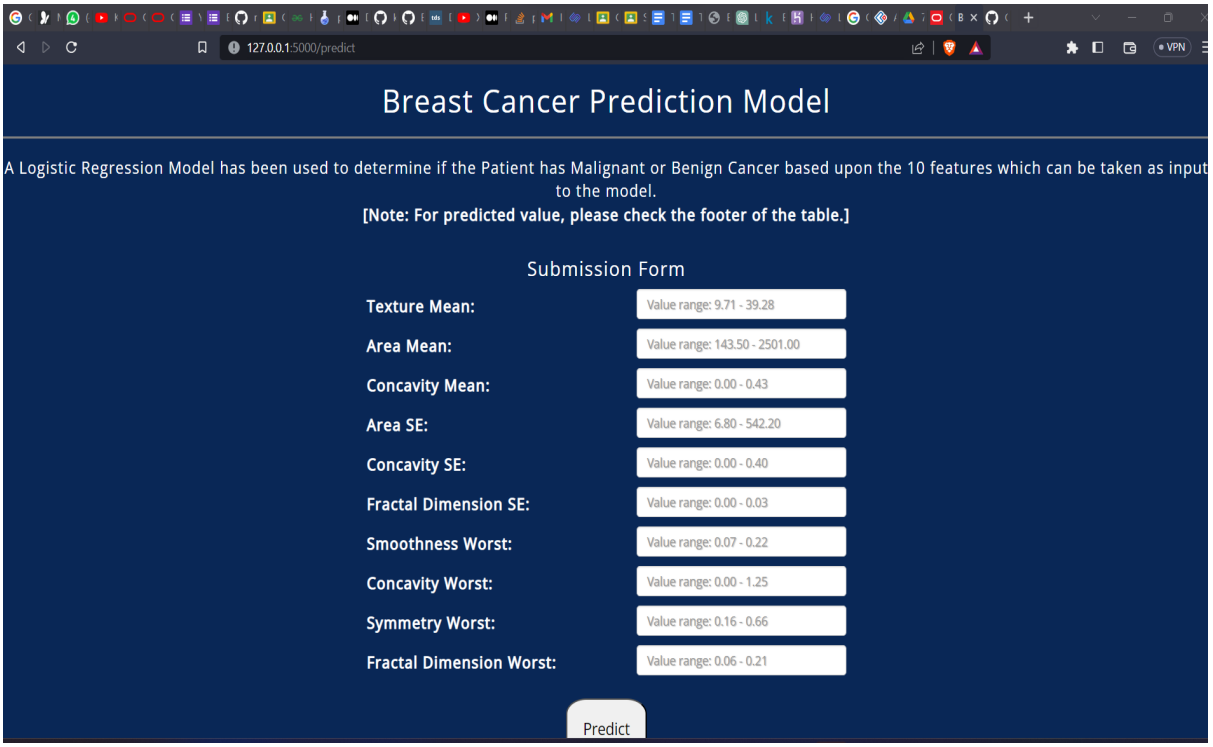
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave _points_mean	...	radius_worst	tex
0	842302	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	25.38	
1	842517	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	24.99	
2	84300903	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	23.57	
3	84348301	1	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	14.91	
4	84358402	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	22.54	

5 rows × 32 columns

```
# turning into categorical data
data['diagnosis'] = data['diagnosis'].astype('category',copy=False)
plot = data['diagnosis'].value_counts().plot(kind='bar', title="Class distributions \n(0: Benign | 1: Malignant)")
fig = plot.get_figure()
```


breast cancer model.ipynb M
actual_model.ipynb U
breast cancer.csv U
app.py U
style.css U
index.html result\... U
result.html U
breast cancer model.ipynb > Importing all the libraries > data.isnull()
Code
Markdown
Run All
Clear All Outputs
Restart
Variables
Outline
Python
#normalising the train-split more
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
X_scaled
[71]
0.0s
array([[-0.23640517, 1.09706398, -2.07333501, ..., 2.29607613,
2.75062224, 1.93701461],
[-0.23640344, 1.82982061, -0.35363241, ..., 1.0870843 ,
-0.24388967, 0.28118999],
[0.43174109, 1.57988811, 0.45618695, ..., 1.95500035,
1.152255 , 0.20139121],
...,
[-0.23572747, 0.70228425, 2.0455738 , ..., 0.41406869,
-1.10454895, -0.31840916],
[-0.23572517, 1.83834103, 2.33645719, ..., 2.28998549,
1.91908301, 2.21963528],
[-0.24240586, -1.80840125, 1.22179204, ..., -1.74506282,
-0.04813821, -0.75120669]])
#splitting the data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.30, random_state=42)
[72]
0.1s
#train the model
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train, y_train)
y_pred = lr.predict(X_test)


breast cancer model.ipynb M
actual_model.ipynb U
breast cancer.csv U
app.py U
style.css U
index.html result\... U
result.html U
breast cancer model.ipynb > Importing all the libraries > data.isnull()
Code
Markdown
Run All
Clear All Outputs
Restart
Variables
Outline
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy:.2f}')
[76]
0.0s
Accuracy: 0.98
import pickle
pickle.dump(lr, open('model.pkl','wb'))
[77]
0.0s
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import classification_report, confusion_matrix
model=GaussianNB()
model.fit(X_train,y_train)
nb_pred=model.predict(X_test)
print(classification_report(y_test, nb_pred))
accuracy = accuracy_score(y_test, nb_pred)
print(f'Accuracy: {accuracy:.2f}')
[78]
0.0s
precision recall f1-score support
0 0.94 0.95 0.95 108
1 0.92 0.90 0.91 63
accuracy 0.94 171
macro avg 0.93 0.93 0.93 171
weighted avg 0.94 0.94 0.94 171



Conclusions

- The machine learning model developed using the "Breast Cancer Data" can accurately predict the likelihood of cancer in patients.
- The model's performance can be further improved by incorporating additional features or by using more advanced machine learning algorithms.
- We have further implemented the model using flask, to help us give dynamic input and get the possible prediction for the same.

Future Work

- It is possible to explore the use of more advanced machine learning techniques such as deep learning or ensemble models to improve the accuracy of the predictions.
- The model can be integrated with existing healthcare systems to aid in cancer detection, diagnosis and improve patient outcomes.
- Finally, it is important to continuously monitor the model's performance and update it periodically to reflect new data and changes in patient demographics or risk factors.

References

1. Applications of Machine Learning in Cancer Prediction and Prognosis - (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675494/#:~:text=Among%20the%20better%20designed%20and,cancer%20susceptibility%2C%20recurrence%20and%20mortality)
2. Machine learning applications in cancer prognosis and prediction - (https://www.sciencedirect.com/science/article/pii/S2001037014000464)
3. Cancer Prediction using Machine Learning - (https://ieeexplore.ieee.org/document/9754059)
4. ML Systems in the Real World: Cancer Prediction - (https://developers.google.com/machine-learning/crash-course/cancer-prediction)
5. Linear Regression in Python (https://realpython.com/linear-regression-in-python/)
6. Scikit Learn Documentation (https://scikit-learn.org/stable/index.html)
7. Matplotlib- Visualisation with Python (https://matplotlib.org/)
8. Understanding Cancer using Machine Learning (https://towardsdatascience.com/understanding-cancer-using-machine-learning-84087258ee18)
9. Cancer Detection Github Projects (https://github.com/topics/cancer-detection)

10. Predicting Cancer- Regression (https://www.kaggle.com/code/jagannathrk/predicting-breast-cancer-logistic-regression)
11. Machine Learning with Python (https://www.geeksforgeeks.org/machine-learning-with-python/)
12. Python Machine Learning Course (https://www.youtube.com/watch?v=7eh4d6sabA0)
13. PyImage Machine Learning (https://pyimagesearch.com/2019/01/14/machine-learning-in-python/)
14. Kim, H. Y., Shim, H. S., & Kim, L. (2018). Machine learning-based prediction of cancer survival
15. https://www.reddit.com/r/cbirt/comments/1206v7u/machine_learningaided_multiscale_transcriptomics/
16. https://www.kaggle.com/datasets/denizkavi1/brain-tumor
17. https://www.kaggle.com/datasets/rishidamarla/cancer-patients-data?resource=download
18. https://nuadox.com/post/7124313746745589 76/pediatrics-cancer-atlas-ai
19. https://link.springer.com/article/10.1007/s00259-023-06145-z
20. https://scienmag.com/new-machine-learning-model-improves-prediction-of-prostate-cancer-recurrence/?feed_id=66522&_unique_id=63ebd871cb5f5