Breast cancer prediction

Group 7

Group members

Name	Email	Role
Vrushabh Ajaybhai Desai	VrushabhAjaybhaiDesai@my.unt.edu	Document,Code
Varsha Salil	VarshaSalil@my.unt.edu	Document, Code
Kareem Baba Shaik	kareembabashaik@my.unt.edu	Code, Testing
Manasa Kilaru	manasakilaru@my.unt.edu	Dataset, Code
Alekhya Sree Jetti	AlekhyaSreeJetti@my.unt.edu	Document, Testing

Collaboration

- Thirty minute in-person meetings on Fridays after class.
- Communication via Email, Canvas, WhatsApp groups and discord.
- Collaborated over Zoom Meetings over the week.

Abstract

- Breast cancer is one of the leading causes of death in women these days. The condition being gradually progressing, will exhibit most of the symptoms during more advanced stages. The severity of the condition can be kept under control by early prediction and proper treatment.
- The current methods for diagnosis include physically invasive tests and chemical tests on bloodwork. Body cells are then classified into Benign (or non-cancerous) cells or Malignant(cancerous) cells. Cancerous cells can also be detected by analyzing several body parameters like clump thickness, cell size, cell shape etc.
- In this project, we propose a non-invasive study and classification of a patient’s body parameters of interest to classify breast cells into two categories i.e, Benign or Malignant, depending on the patient’s diagnosis result.

What is Different?

- Kaggle includes various codes where they have included models like CNN and random forest.
- In this case, we are attempting to achieve the same goal using a Naive Bayes model and different preprocessing methods.

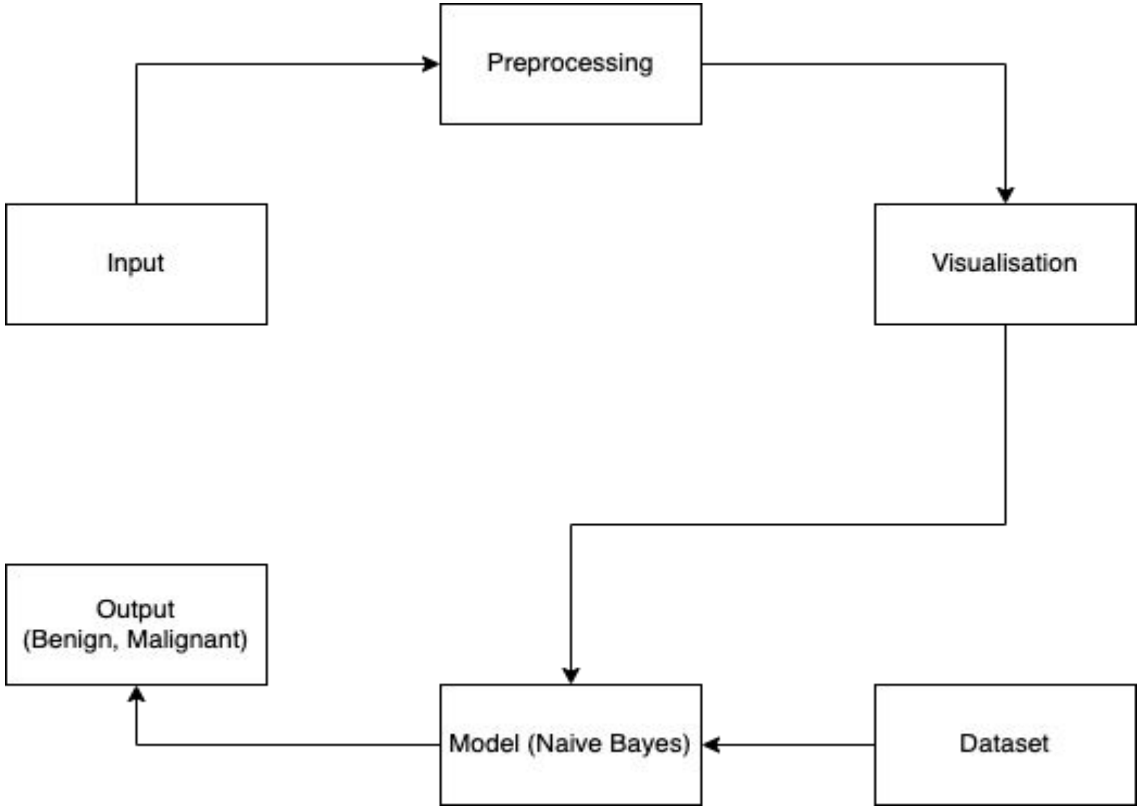
Project Design

- The main objective of this project is to diagnose Breast cancer in women by monitoring various body parameters that are known to indicate the presence of the condition. For this, we are using a Gaussian Naïve Bayes model to classify the dataset. A Gaussian Naïve Bayes model is a supervised, probabilistic Machine learning model used for classification purposes.
- The project will be coded in Python using Jupyter notebook.

Workflow

- The first step includes obtaining the relevant data for the diagnosis and data preprocessing for classification, by removing null values, dropping unnecessary columns etc. The following dataset will be used for training and testing purposes: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/.
- The next step is data visualization.
- A Gaussian Naïve Bayes model is generated next from sklearn.naive_bayes package and is fit with the data for training purpose.
- The model then classifies the data into Benign or malignant and returns the predication along with the accuracy of the prediction.

Flow Chart


Preprocessing
Input
Visualisation
Output
(Benign, Malignant)
Model (Naive Bayes)
Dataset

Milestones

- Project proposal
- Algorithm generation
- Code generation
- Testing the code for efficiency

Future Scope

- This project can be further expanded to diagnose other health conditions when the relevant data is available.
- Multiple Machine learning models can be used to compare and improve the diagnosis.
- Implement validation set to tweak the hyper parameters and check correct value for the model.

Advantages

- Economic and faster diagnosis of Breast cancer.
- Non-invasive diagnosis using Gaussian Naive Bayes model.

References

- https://ieeexplore-ieee-org.libproxy.library.unt.edu/document/9754059/ : The link to the base paper for the project.
- https://www.kaggle.com/code/buddhiniw/breast-cancer-prediction/notebook : Reference for dataset, implementation and testing