



CISC 482

COMPUTER SYSTEMS SEMINAR

Breast Cancer Classification Model using Machine Learning

Author:
Hrishikesh Lamichhane

Student number:
842963

Sunday 28th April, 2024

1 Introduction

Breast cancer is one of the most prevalent diseases affecting women worldwide. As a public health major with a passion for chronic disease epidemiology, I am interested in understanding this illness, through the screening point of view. Thus, I am building a classification machine learning project focused on predicting the nature of breast cancer cells—whether they are benign or malignant.

Utilizing a dataset with attributes such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, alongside the diagnosis (benign or malignant) as labels, this project aims to develop a predictive model. By applying advanced machine learning algorithms, this model will analyze patterns to distinguish and indicate the the cancer's nature.

This project directly aligns with my core curriculum and personal interest. Through courses like Bio science I, I have delved into the intricacies of cancer biology, understanding its various types and underlying mechanisms. Additionally, my hands-on experience in bio science labs has provided practical insights into cancer genetics and cellular behavior. Taking Health Assessment class have allowed me to understand better the importance of screening diseases with reliability and validity and linear algebra and Biostatistics have allowed me to grasp the mathematical approach to this machine learning project.

By integrating my academic background with my passion for public health and epidemiology, this project embodies a convergence of knowledge and purpose. Ultimately, the predictive model developed through this research has the potential to contribute significantly to early detection efforts and personalized treatment strategies for breast cancer—a cause I am deeply committed to advancing.

1.1 Data Description

The data was retrieved from Kaggle. The data is from The Breast Cancer (Wisconsin) Diagnosis Dataset. The study has total 32 attributes and 569 entries. The data contains information on different attributes after cellular and tissue level assessment using the Fine-Needle Aspiration (FNA) which is a minor surgical procedure.

The attribute comprises of ten real valued features, including radius, texture, perimeter, area, smoothness, compactness etc from the sample cell analysis from the lump or the mass investigation of the cancerous tissue. Each features has its values and one of the columns contain data coded as 'Benign' and 'Malignant' expressed in the binary (0 and 1).

1.2 Data Description

The dataset represents samples of pregnant women (Fig 6), including characteristics such as age, blood pressure, glucose levels, heart rate, and body temperature, with the aim to predict the risk level of maternal health.

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

5 rows x 32 columns

texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2854	0.4601	0.11890
23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902
25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758
26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300
16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678

Figure 1: Data Head

2 Exploratory Data Analysis

My exploratory data analysis includes visualizing the overall data in the beginning followed by visualizing graphs and plots for the features for benign group and malignant groups separately. This gives me the idea of whether or not machine learning will efficiently classify the cancers in a distinct way. I have used seaborn and matplotlib package for data exploration.

2.1 Data Comparison

Firstly, the data was described to display the basic statistics of all the necessary features.

2.2 Response Variable (Target Variable)

Our target variable is the feature 'diagnosis'. The target variable is characterized into two groups- benign or malignant cancer. M indicates Malignant and B indicates Benign form of cancer. For label encoding, the target variable is expressed in binary. Benign is '0' and Malignant is '1'.

2.3 Predictor Variables (Features)

My predictor variables are all the features present in the dataset. Since,

Figure 1 includes including mean, standard deviation, minimum value, maximum value and quarterlies. The data is described for all the features.

	id	radius_mean	texture_mean	perimeter_mean	area_mean	\
count	1.000000e+02	100.000000	100.000000	100.000000	100.000000	
mean	1.547093e+07	14.707780	19.692200	96.471200	703.293000	
std	3.066549e+07	3.349245	3.759176	23.187471	320.152301	
min	8.571500e+04	8.196000	10.380000	51.710000	201.900000	
25%	8.542642e+05	12.457500	16.760000	82.270000	476.800000	
50%	8.593735e+05	14.335000	20.190000	94.365000	643.650000	
75%	8.610460e+06	17.155000	22.150000	114.400000	916.875000	
max	8.613550e+07	25.220000	27.540000	171.500000	1878.000000	

	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\
count	100.000000	100.000000	100.000000	100.000000	
mean	0.102040	0.126576	0.114752	0.063801	
std	0.013151	0.061123	0.078862	0.038104	
min	0.073550	0.037660	0.000692	0.004167	
25%	0.093402	0.080118	0.041847	0.029170	
50%	0.101150	0.118150	0.107150	0.065810	
75%	0.110375	0.156800	0.166350	0.087590	
max	0.142500	0.345400	0.375400	0.184500	

Figure 2: Data Describe

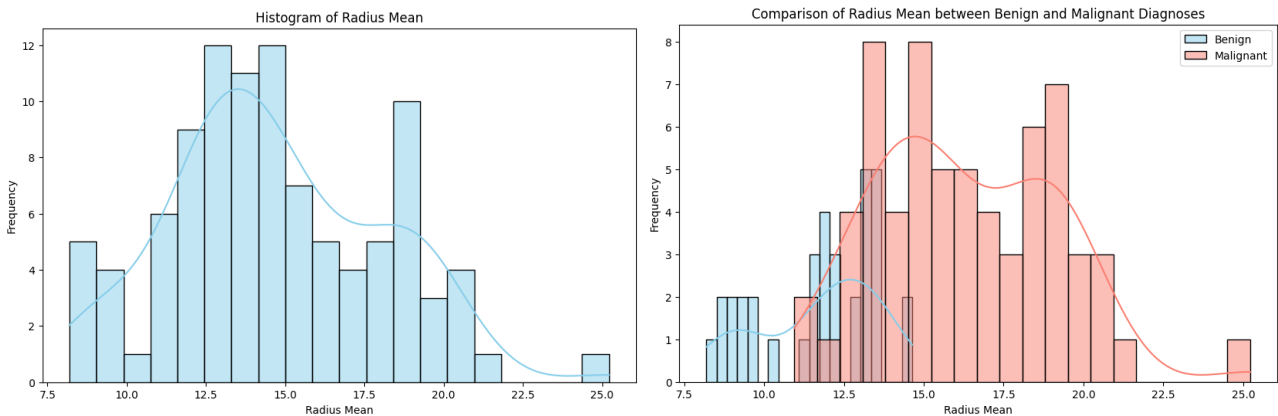


Figure 3: Histogram comparison

The following figures are the histplot of all the predictor variables against the response variable RiskLevel. We can see how data is distributed in three categories: Low risk(0), Mid risk(1) and High risk(2). From the visualization, we can see that the women with higher SystolicBP, DiastolicBP and BS have higher risk level.

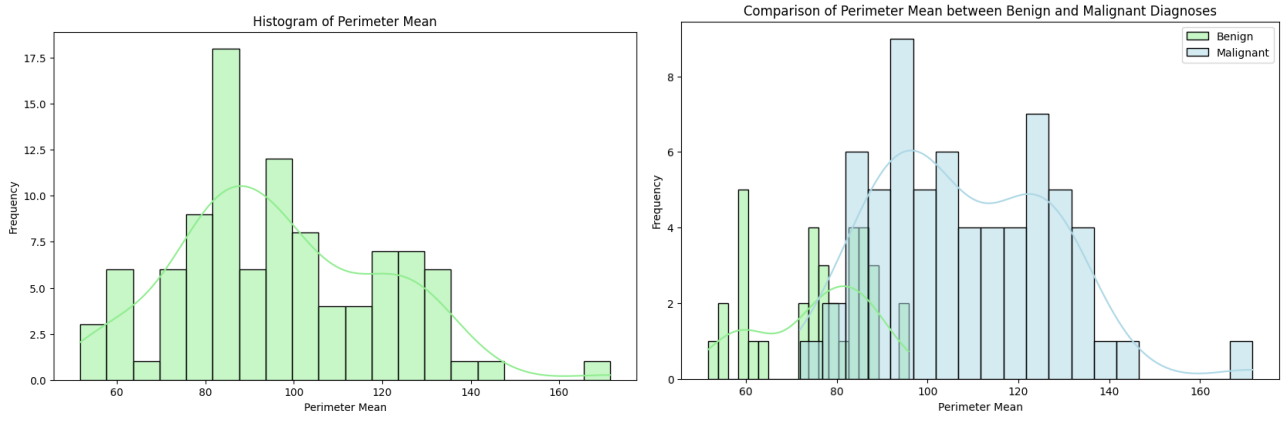


Figure 4: Histogram comparison

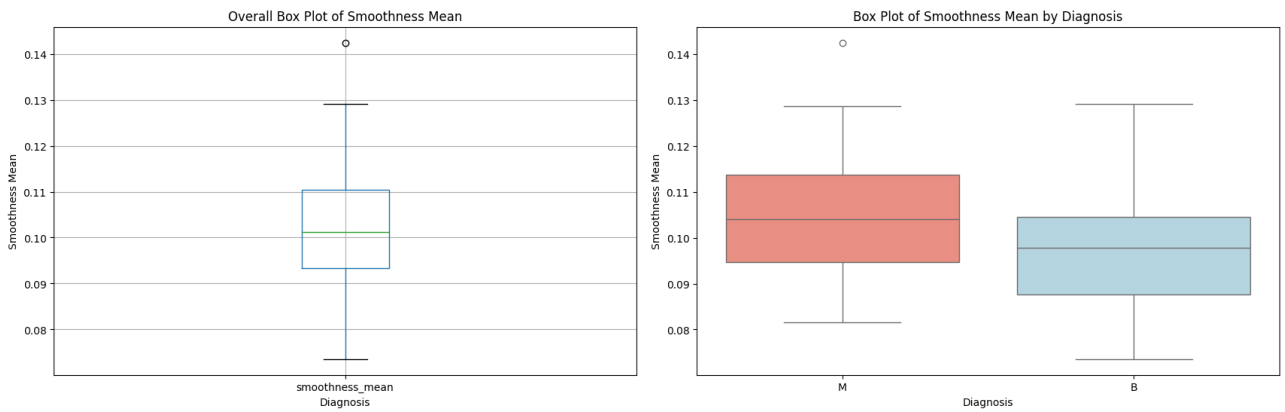


Figure 5: Box plot comparison

2.4 Data Cleaning

Data cleaning was not necessary for my dataset because there were no null values present in my data. There were few outliers, but the number of outliers were negligible hence, I did not alter them.

2.5 Data Splitting

I splitted the 80 percent of total data into training data and the rest of the 20 percent for testing purposes. The X variable was Features which is independent and Y was Labels which is dependent in our models. The shuffle is set to be true and the random state is set to be 0. This helped us generalize the given data.

2.6 Data Scaling

The 'standardscaler' class was imported 'sklearn.preprocessing'. A list was created which consisted of the name of all the features. The 'fit-transform' method was applied to the given columns and scaled in such a way that the numerical values were standardized subtracting by the mean and dividing by the standard deviation.

3 Data Modeling and Analysis

3.1 Logistic Regression Model

Our dependent variable is Cancer-Type in our data. It is the categorical variable so I used logistic regression. Theoretically, the sigmoid function transforms the combination of independent variables into the probabilities, which will eventually make predictions. The multi-class classification model was utilized using the 'sklear.linear-model' module and 5 folds of cross-validation were performed.

3.2 Decision Tree Classification Model

The decision tree classifier 'dtree' was initialized using the 'DecisionTreeClassifier' with the use of 'sklearn.tree' module. The cross validation with five folds was performed resulting in five cross validation scores followed by calculating the mean of all the cross validation scores.

3.3 Neural Network Model

The first layer had 64 neurons, followed by 32 neurons in the second layer followed by 1 neuron with sigmoid activation function. The rate of dropout during transition of each layer was 0.5. The main purpose of this was to prevent overfitting. The model was compiled using Adam Optimizer with the learning rate of 0.001. The total number of epochs was set at 50, with the batch size of 32. The test loss and test accuracy were noted and analyzed.

3.4 KNN Model

The KNeighborsClassifier class from sklearn.neighbors was imported and the number of neighbors was set to 6 and the model was trained using fit method while the testing was achieved using the 'predict' method. Accuracy score was generated to evaluate the performance of the model we built.

4 Discussion

The cross validation scores in the logistic regression model averages to 0.93 which depicts that the performance was consistent throughout the five folds of cross validation and rather very strong. The test accuracy was calculated to be 95 percent which demonstrates high accuracy. For decision tree classifier, the cross validation scores range from 0.81 to 0.93 with an average of 0.9, which suggests that the the decision tree model predicts correctly 90 percent of the time. The KNN model suggests 95 percent accuracy so it proves a strong model to generalize in industry use. For Neural Network model, the test loss was calculated to be 0.205 with the test accuracy of 94.99 percent. This reveals that the neural network model is predicting the data with relatively low loss and high accuracy at the same time.

Model	Cross Validation Score	Mean CV Score	Test Accuracy
Logistic Regression	[0.9375, 0.875, 1.0, 0.9375, 0.9375]	0.9375	95.0%
Decision Tree Classifier	[0.875, 0.8125, 0.9375, 0.9375, 0.9375]	0.9125	-
KNN Model	-	-	95.00%
Neural Network	-	-	94.99%

Figure 6: ML Models and Accuracy

5 Recommendation

In future work, I will be expanding the data and taking a bigger sample size. I will look into validating ML for lesser features by analyzing correlation using Heat Map and picking minimal features. It is recommended to use the machine learning models and concepts to design the model for other type of cancer like ovarian cancer. More work could be done to classify even further and more concise/stages of malignant cancer. More machine learning models could be studies comparing the accuracy of each.