

Breast Cancer Detection Model Using Logistic Regression

Author

Perikles Tsikrikis

Name

Abstract

Due to the increasing prevalence of breast cancer among females, it is becoming increasingly important to develop and improve methods of detection, as early there is a positive correlation between detection and survival rate. Machine Learning is one such tool that is being used to great effect. The data set selected to train and test our model from the Wisconsin breast cancer database made public through kaggle. The set is composed of 570 records, 10 primary features and 20 subsequently derived features each pair corresponding with a primary feature. We selected a Logistic Regression model as several medical journals cited Logistic Regression as favorable in breast cancer classification problems . Our model takes input data on 21 tumor features (stripped down from 30) and using a 70/30 train test split is able to accurately classify with 99.1% accuracy whether a given tumor is Benign or Malignant.



Introduction

According to data published by the Canadian Cancer Society, a leading authority on cancer research and statistics. Breast cancer is the most common type of cancer developed among women and is the second leading cause of cancer related death. With cancer rates steadily increasing and the growing number of cancer related deaths, there has never been a greater need for the improvements in the methods of cancer treatment and prevention. Cancer specialists agree, the current most effective method of cancer treatment and prevention is early detection. Fortunately alongside the increase in cancer cases, has come great technological breakthroughs in the field of Artificial Intelligence and Machine Learning. Machine learning, a subset of Artificial Intelligence, uses mathematical models combined with input data in which models are trained in order to learn and improve on ability to perform a specified task. In the case of breast cancer tumors using breast cancer data sets, models can be trained to classify breast cancer tumors, more efficiently and accurately than current medical professionals. The reduced risk of human error and increase in speed and efficiency is what drove us to apply our knowledge of Machine Learning in order to develop a model to facilitate the improvement of breast cancer detection and reduce over breast cancer fatality.



Project Description

When detecting breast cancer the first and earliest indicator is the formation of a tumor. Breast cancer tumors have two primary classifications, Malignant and Benign. Benign tumors are tumors in which growth is slow and remain in their present location confined to the tumor walls. They seldom have the ability to spread through the body. These tumors are easily treated and non problematic. Malignant tumors however pose a serious risk to the wellbeing of the patient as they grow at an uncontrollable rate, spreading to different organs through the body making it difficult to eliminate as stages progress. The goal of our project is to accurately and efficiently predict the classification of breast cancer tumors using supervised learning that takes quantitative tumor data and resulting classification in order to identify and prevent Malignant tumors from spreading and causing further harm to the patient.




In order to accurately classify a given breast cancer tumor we must first obtain as much relevant quantitative data on the attributes of the tumor as possible. We spent a great deal of time carefully selecting a data set with the highest possible quality data that was both accurate and complete as the trainability and success of a machine learning model heavily depends on the type and quality of data used. The data must be accurate, complete and of sufficient size in order to train the model with a high degree of accuracy. The attributes of the tumor in the context of Machine learning are called features. Features are used as the benchmark for establishing correlations between feature values and probability of classification with respect to the target class.

After thoroughly researching the many current methods of machine learning models used for breast cancer detection, the consensus of many studies such as a study published by the Electrochemical society and a peer review journal published on the European Chemical Bulletin were in agreement with Logistic Regression being one of the primary and best performing models for tumor classification.

When formulating the initial proposal for the project, several key questions arose. We knew the project relies on the availability of historical breast cancer tumor data from medical databases. Some initial concerns were Limited access to comprehensive datasets, the accuracy and completeness of the tumor data, missing or noisy features and low-quality data. The choice of machine learning algorithms and model complexity was also considered as it directly affects the interpretability and explainability of the predictions.



Methodology



We selected the dataset “Breast Cancer Wisconsin (Diagnostic)” from Wisconsin Breast Cancer Data Base, made public by Kaggle, an online Computer science and machine learning resource. We chose this set because of its robust feature offering and large number of instances. The features in this data set were derived from the digital imaging of fine needle breast mass samples. Images were then analyzed to produce and quantitate cell nuclei attributes. Initial features of the data set included ID number, Diagnosis (M = malignant, B = benign) and ten numeric features of each cell nucleus including: Radius calculated by (mean of distances from center to point on perimeter), Texture calculated by (standard deviation of gray-scale values), Perimeter, Area, Smoothness calculated by (local variation in radius lengths), Compactness calculated by $(\text{perimeter}^2 / \text{area} - 1.0)$, Concavity, Concave points, Symmetry and Fractal dimension. Along with these 10 primary features included were 20 subsequent features derived from their primary counterparts (two for each primary feature). The subsequent features were: Standard error and “Worst” (defined as “the mean of the three largest values”). These subsequent “Worst” features are incredibly interesting and useful as they represent a hybrid concentration of their primary counterpart and help increase the depth of the model and improve training accuracy through correlation with their primary feature in relation to the target.

In order to perform Logistic regression on the given feature data, we first took inventory of and analyzed the data values, primarily checking for missing

values and differing data types. Our initial scan informed us of the completeness of the data. "Id" was identified as unimportant and our target class was in need of type conversion. After adjusting the target class from character value to numeric, where (M = 1, B = 0). We were able to begin the correlation evaluation process.

Feature-Target correlation was first analyzed in order to identify features with low correlation that may impede prediction accuracy in the training process and slow down run time.

Highest-Target Feature Correlation:

- Perimeter_worst
- Radius_worst
- Concave points_mean
- Perimeter_mean
- Radius_mean

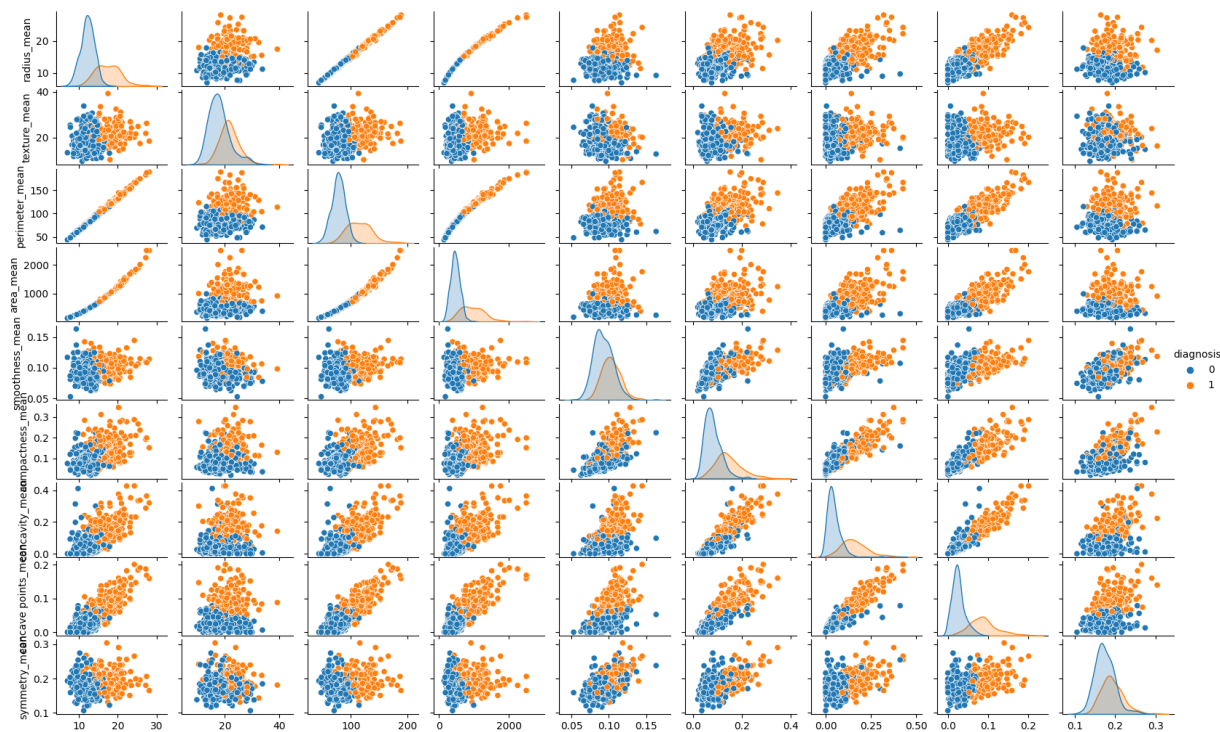
Poorest-Target Feature Correlation:

- Symmetry_se
- Texture_se
- Fractal dimension_se
- Fractal dimension_mean

diagnosis	1.000000
radius_mean	0.730029
texture_mean	0.415185
perimeter_mean	0.742636
area_mean	0.708984
smoothness_mean	0.358560
compactness_mean	0.596534
concavity_mean	0.696360
concave points_mean	0.776614
symmetry_mean	0.330499
fractal_dimension_mean	-0.012838
radius_se	0.567134
texture_se	-0.008303
perimeter_se	0.556141
area_se	0.548236
smoothness_se	-0.067016
compactness_se	0.292999
concavity_se	0.253730
concave points_se	0.408042
symmetry_se	-0.006522
fractal_dimension_se	0.077972
radius_worst	0.776454
texture_worst	0.456903
perimeter_worst	0.782914
area_worst	0.733825
smoothness_worst	0.421465
compactness_worst	0.590998
concavity_worst	0.659610
concave points_worst	0.793566
symmetry_worst	0.416294
fractal_dimension_worst	0.323872
Name: diagnosis, dtype: float64	

Following Correlation analysis, and given our feature surplus we then removed all features with correlation to the target class less than 0.4.

Furthermore we analyzed Correlation between features. Below is the feature correlation matrix from features 2-10 (excluding target feature 'Diagnosis')



We deduced several things from this matrix representation.

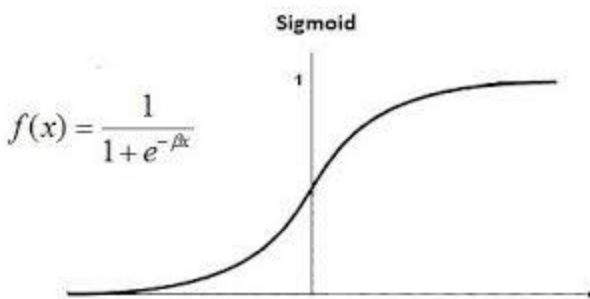
The features are a function of the standard distribution. As evident by the correlation plot, the features with the highest correlations were (perimeter_mean and radius_mean), (area_mean and radius_mean) and (area_mean and perimeter_mean). This is an expected correlation as these features are a variety of measurements relating to tumor size. As we can conclude from these inter feature correlations, as well as their correlations with the target. There is a clear relationship between the size of the tumor and probability of malignancy.

FINAL TARGET CORRELATIONS:

```
diagnosis          1.000000
radius_mean        0.730029
perimeter_mean     0.742636
area_mean          0.708984
compactness_mean   0.596534
concavity_mean     0.696360
concave points_mean 0.776614
radius_se          0.567134
perimeter_se       0.556141
area_se            0.548236
concave points_se  0.408042
radius_worst       0.776454
texture_worst      0.456903
perimeter_worst    0.782914
area_worst         0.733825
smoothness_worst   0.421465
compactness_worst  0.590998
concavity_worst    0.659610
concave points_worst 0.793566
symmetry_worst     0.416294
Name: diagnosis, dtype: float64
```


After cleaning up cleaning the data and eliminating unnecessary features, the data standardization process took place where we standardized the model Features in a dataset may have different scales, units, or measurement ranges. Some features might have values in the thousands, while others have values between 0 and 1. Without standardization, features with larger scales can dominate the learning process, leading to biased and inefficient learning. Standardized features make it easier to interpret the importance and impact of each feature on the model's predictions, as their magnitudes are comparable.

Logistic regression is a form of linear regression used for classification problems, Logistic Regression uses a special function called the "Logistic function" or "Sigmoid function" to make predictions. The sigmoid function takes any real-valued number and narrows them into a range between 0 and 1. The resulting value then represents the probability of an instance classified as Malignant or Benign. Logistic Regression is efficient in classifying unknown records, and performs well for simple data sets where data is linearly separable.



According to several studies from the Electromechanical society and Euro chemical bulletin, a peer reviewed journal indicates that logistic regression is

one of the highest accuracy models used in breast cancer classification problems.

We then explored different variations of training and testing splits, after several different configurations, a 60/40 split yielded the highest level of accuracy. In order to generate consistent results we used the starting point of state 5 as the random state parameter in our Logistic regression function call.

Results

The results of our Logistic Regression Classification model using 60/40 (342/228) train test split were as follows.

We achieved:

Accuracy

Training: 97.65 %

Testing: 99.12 %

Precision

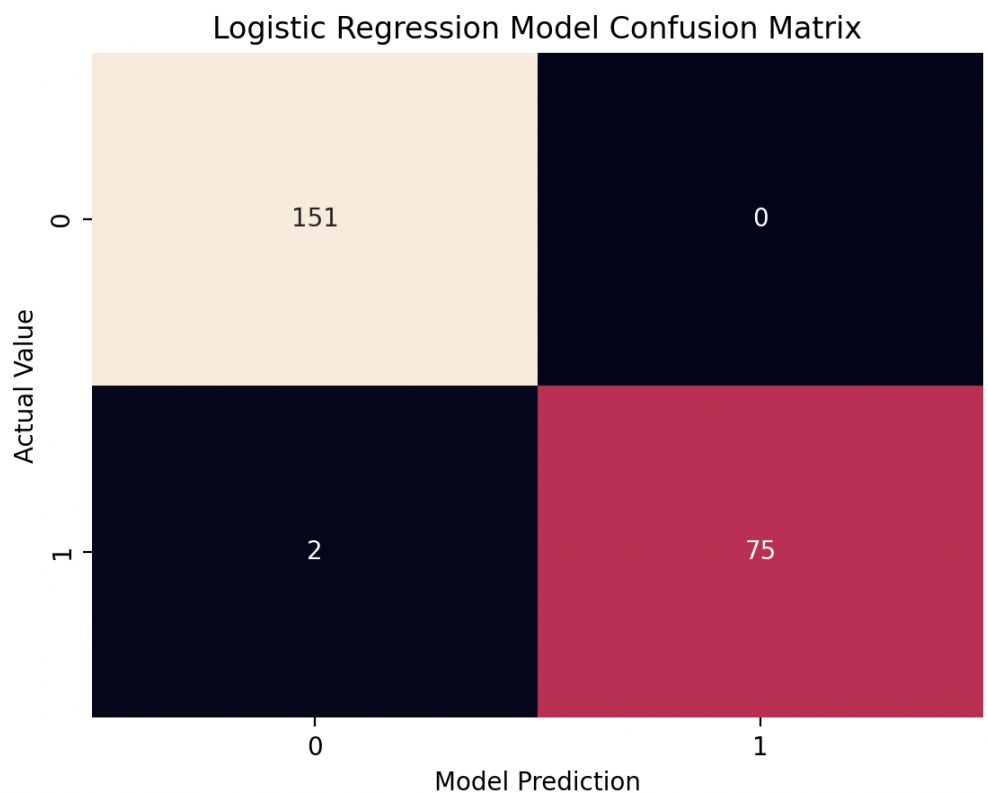
Training: 97.74 %

Testing: 100.0 %

Recall


Training: 96.3 %

Testing: 97.4 %



Precision focuses on how many of the positive predictions are actually correct. Recall focuses on how many of the actual positive instances were correctly predicted by the model. We are proud of our models accuracy and are able to say with confidence that the results of future testing data are likely to be achieved with a high accuracy, precision and recall.

Conclusion

 Machine learning using Logistic Regression is an incredibly effective way of analyzing tumor data and effectively predicting and detecting precedence of Malignant Breast Cancer tumors. Since early-detection is one of the most important aspects of treatments concerning the disease, it is evident that Machine Learning is a very useful tool in combating it. Using this model as a tool for early detection will help decrease its progression and prevent it from killing the patient with appropriate treatment measures.

References

<https://cancer.ca/en/cancer-information/cancer-types/breast/statistics#:~:text=Breast%20cancer%20is%20the%20most,from%20cancer%20in%20Canadian%20women>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/#:~:text=Logistic%20regression%20is%20used%20to,the%20observed%20event%20of%20interest.>

<https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/>

