

Breast Cancer Detection

With Support Vector Machine (SVM)

Presented by

David Orizu

Introduction

Globally, breast cancer is the most prevalent cancer among women and ranks second in terms of mortality rates. The diagnosis process for breast cancer typically begins when an unusual lump is detected, either through self-examination or an x-ray, or when a small calcium deposit is observed on an x-ray. Once a suspicious lump is identified, a doctor will diagnose it to ascertain if it is Malignant (cancerous) and whether it has metastasized to other parts of the body or is benign (not cancerous).

A benign tumor has distinct, smooth, regular borders. A malignant tumor has irregular borders and grows faster than a benign tumor.

A malignant tumor can also spread to other parts of your body. A benign tumor can become quite large, but it will not invade nearby tissue or spread to other parts of your body.

When a diagnosis is being made, ten real-valued features are computed for each cell nucleus:

- a) radius (mean of distances from the center to points on the perimeter)
- b) texture (standard deviation of gray-scale values)
- c) perimeter (the measure of the core tumor)
- d) area (the measure of the breast tissue)
- e) smoothness (local variation in radius lengths)
- f) compactness ($\text{perimeter}^2 / \text{area} - 1.0$)
- g) concavity (severity of concave portions of the contour)
- h) concave points (number of concave portions of the contour)
- i) symmetry (how much of a proportion difference there is)
- j) fractal dimension ("coastline approximation" - 1)

The purpose of the created model is to detect accurately what type of tumor is found in the breast with the given features. The model will predict if the tumor is Malignant (cancerous).

Approach

After some research and tests, I have chosen to use a Support Vector Machine (SVM) model for this task. The SVM algorithm classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space. I thought of using the SVM algorithm because:

- SVMs perform better with high-dimensional data and are less prone to overfitting compared to decision trees.
- SVMs tend to perform better than Naive Bayes when the data is not linearly separable.

The accuracy of the model is going to be important but the most important metric I will be looking at is the Recall because I don't want the model missing a malignant tumor.

Feature Exploration

Figure 1 is a heatmap for the relationship between all the variables. We will only look at the diagnosis line because that is what is trying to be predicted. The darker red the better relationship.

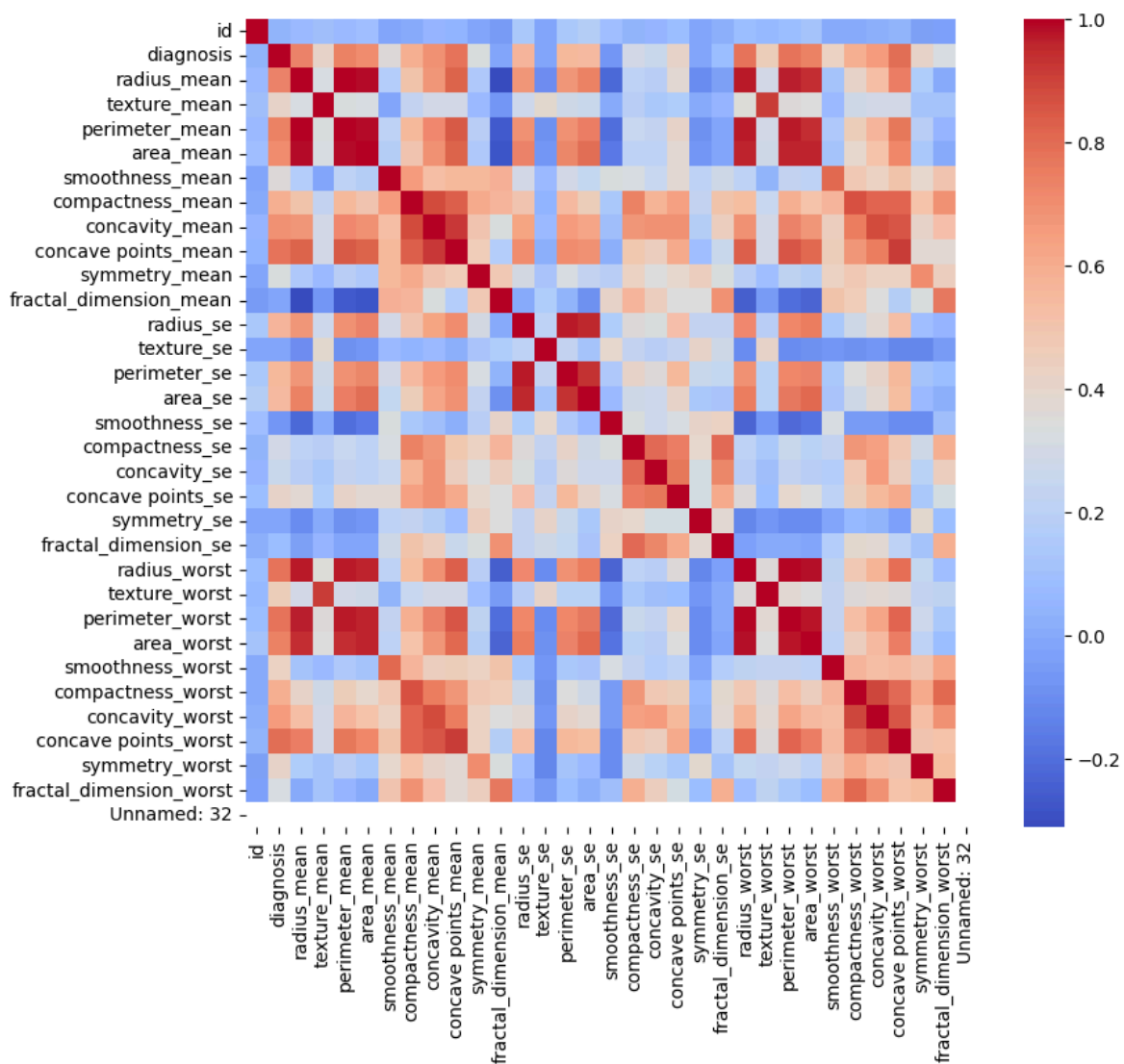


Figure 1

Model performance

Model comparison:

I ran some other models side-by-side with my SVM model to see if any would be better. I compared the model results in Figure 2 & Figure 3.

Accuracy:

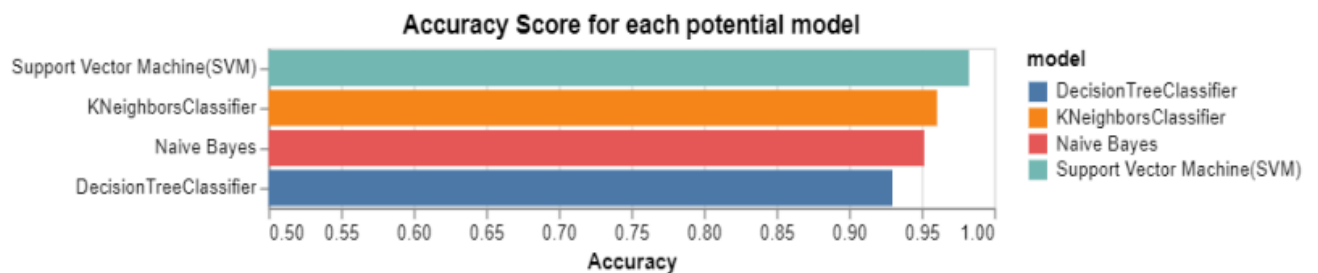


Figure 2

The graph above (Figure 2) shows the accuracy of each model. The SVM model has the highest accuracy with 0.98, while the decision tree has the lowest 0.92

Recall:

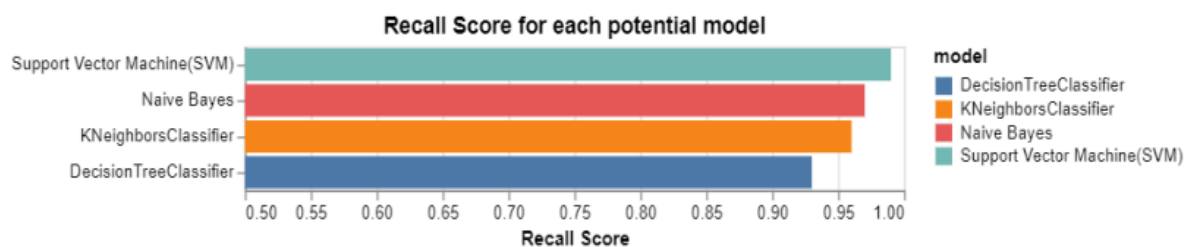


Figure 3

The graph above (Figure 3) shows the recall of each model. The SVM model has the highest accuracy with 0.99, while the decision tree has the lowest 0.93.

Figures 2 and 3 show how much improvement the SVM model is compared to the other classification models.

Let's look more into how good the SVM model did:

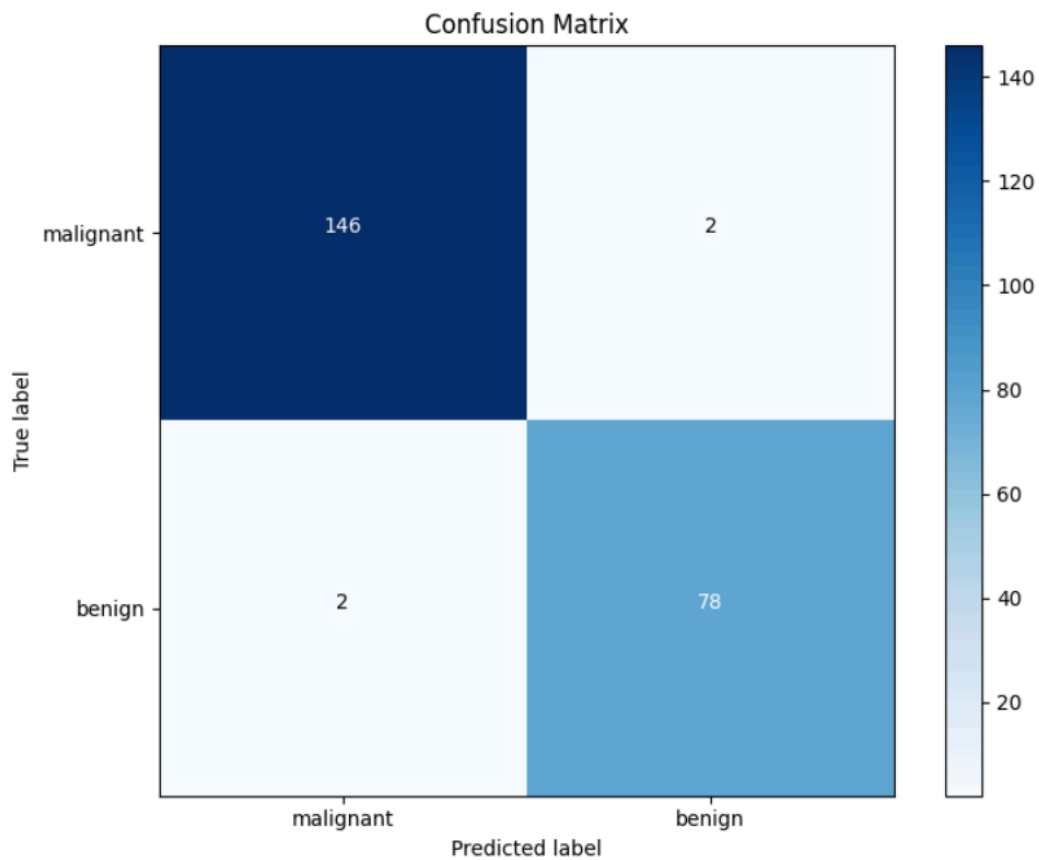


Figure 4

Figure 4 shows the confusion matrix for the results of the model. The dark blue is the number of malignant diagnoses that the model got right, the light blue is the number of benign diagnoses the model got right, and the two white boxes are the numbers the model missed.

The model has an accuracy of 0.98 and a recall of 0.99 for the malignant.

Identifying interesting Finds

I was able to extract the features that were most important to the model's performance.

Figure 5 below shows this:

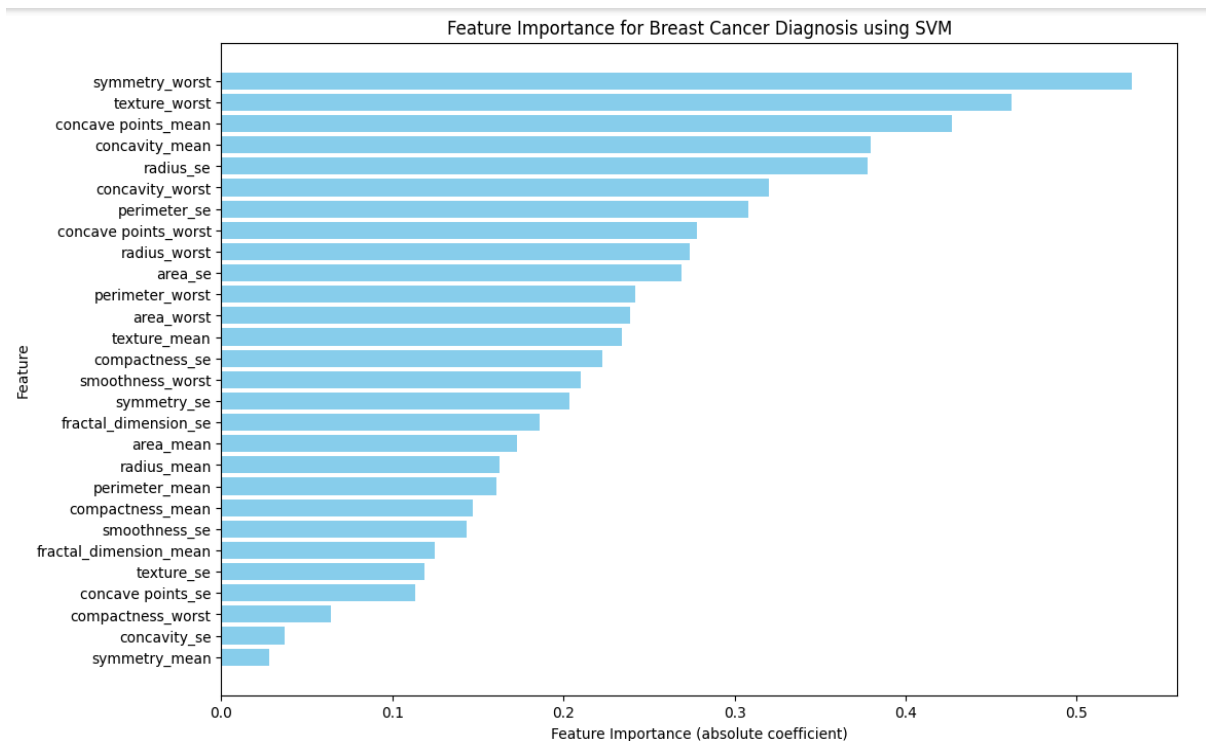


Figure 5

Figure 5 shows the features that we can say would be the most needed in making the model's output accurate. We see that the top 3 most important features are `symmetry_worst` ("Worst" or largest mean value of symmetry of cell nuclei among the features), `texture_worst` ("Worst" or largest mean value of the texture of cell nuclei among the features), and `concave points_mean` (Mean of the number of concave portions of the contour in cell nuclei).

Showing the benefits of technology for the company


Without the help of predictive models, doctors would have to decide what to think of cancer-based on their knowledge of past events, but a doctor would not be able to have as much past data as these models. So, implementing predictive models would help to predict the type of cancer more accurately.

Some other technology has been made; Han et al. reported the identification of benign and malignant masses using CNN for breast ultrasound images, with an accuracy of 90%, and diagnostic sensitivity and specificity of 86% and 96%, respectively. This shows an accuracy for images.

Conclusion

The model that was presented here can predict what type of tumor is in the breast with an accuracy of 98% and a recall of 99%. Implementing this model would help in detecting if a tumor is cancerous early on and this would lead to a better chance of curing or stopping the spread.

Google collab Notebook

 Breast Cancer Detection.ipynb