Breast Cancer Prediction Model and Sister User Interface

Seamus Coyne – coyne.se@northeastern.edu

Northeastern University

DS2500 Intermediate Programming with Data

Prof. Laney Strange

Dec. 5 2023

Breast Cancer Prediction Model and Sister User Interface

Introduction

Breast cancer remains among the leaders in cancer mortality rates, being especially detrimental to women. In fact, it accounts for roughly 30% of new female cancers yearly[1]. With fatality numbers reaching 42,000 women and 500 men per year[2], breast cancer is a common and devastating reality for many families, who oftentimes have no promising avenues to pursue in terms of treatment.

However, many developed nations have found ways to prevent these tragedies. Countries with an abundance in medical resources are often able to employ preventative screen measures, such as annual mammography. These measures ensure early treatment in many cases, contributing to the 4–5-year relative survival rate superseding 80%[3]. Such a figure serves to indicate the importance of rapidity in the diagnosis and treatment of the disease.

Following the determination of a mass from any number of screening methods, the most common next step is to undergo a biopsy. A surgical biopsy involves removing a portion of the unknown mass and viewing it microscopically. Data gathered on the cells can give information on the potential of the cancer to be benign or malignant, allowing physicians and patients to be informed while moving forward. However, depending on medical screening infrastructures in place, this process can be length or entirely impossible. A similar method of analysis may be completed in the form of a fine needle aspiration (FNA). This procedure is often far more available and cost effective in developing nations and underfunded communities and was thus used in this model. To facilitate the rapid diagnosis and treatment of breast cancers, a highly accurate and accessible tool able to be used by both healthcare professionals and patients is needed. As such, this is the final goal of this project.

Dataset Explored

The dataset chosen for this project was retrieved from the publicly available opensource UC Irvine Machine Learning Repository in the .CSV filetype. The set contains 569 entries, 357 of which were ultimately diagnosed as benign with the remaining 212 as malignant. While 32 features were available, only 10 were ultimately used. This was because 2 features were ID and diagnosis columns, and 20 of the remaining 30 were add-ons to the primary data, such as standard error and "worst" measurements. While all 30 features could functionally be used to train the model, doing so ultimately worsened the accuracy of the model and reduced functionality of the user interface.

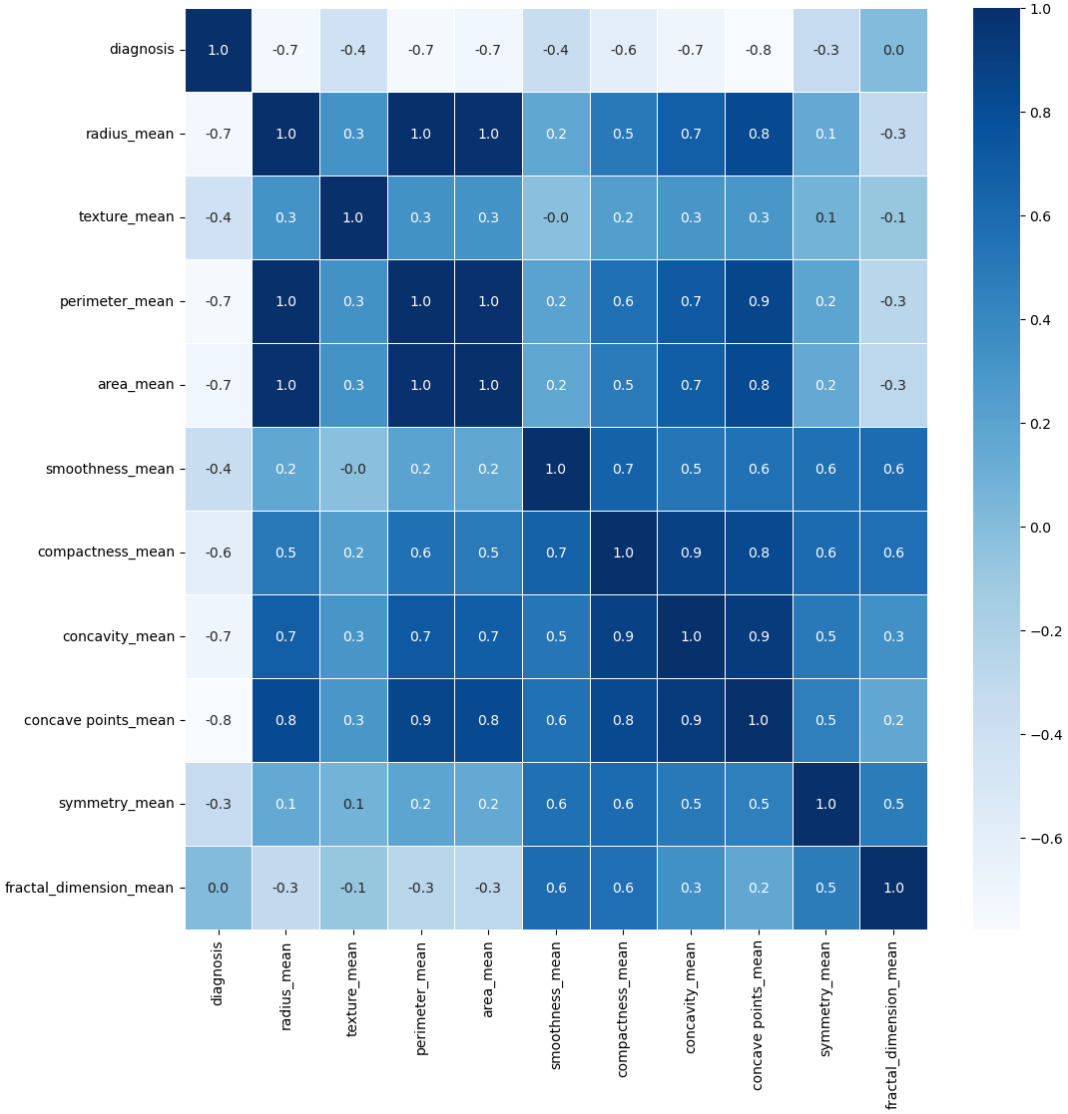
The UCI dataset was selected due to multiple factors, including its lack of patient-identifying markers, numeric datatypes, and usage of ethical data collection methods. All of these factors were important to ensure successful development of the K Nearest Neighbors model. Features retained for training of the model include can be found below:

1. *Radius (mean distance from center to perimeter points)*
2. *Texture (standard deviation of grey scale values)*
3. *Perimeter, area, smoothness (local variation in radius lengths)*
4. *Area*
5. *Smoothness*
6. *Compactness (perimeter2/(area – 1.0))*
7. *Concavity (severity of concave portions of the contour)*
8. *Concave Points (number of concave portions of the contour)*
9. *Symmetry*
10. *Fractal Dimension ("coastline approximation" – 1)*

Methods Used

In order to facilitate further data manipulation and usage, it was first necessary to complete relatively extensive data cleaning and exploration. Work in this stage included dropping unwanted columns from the dataframe and verifying that there were no NaN values (none were found). This portion of work also included generalized data exploration to better-understand what the data showed. An example product can be found at Figure 1, which was used to determine what features would be used to train the model.

Figure I: Heatmap Showing Feature Correlations

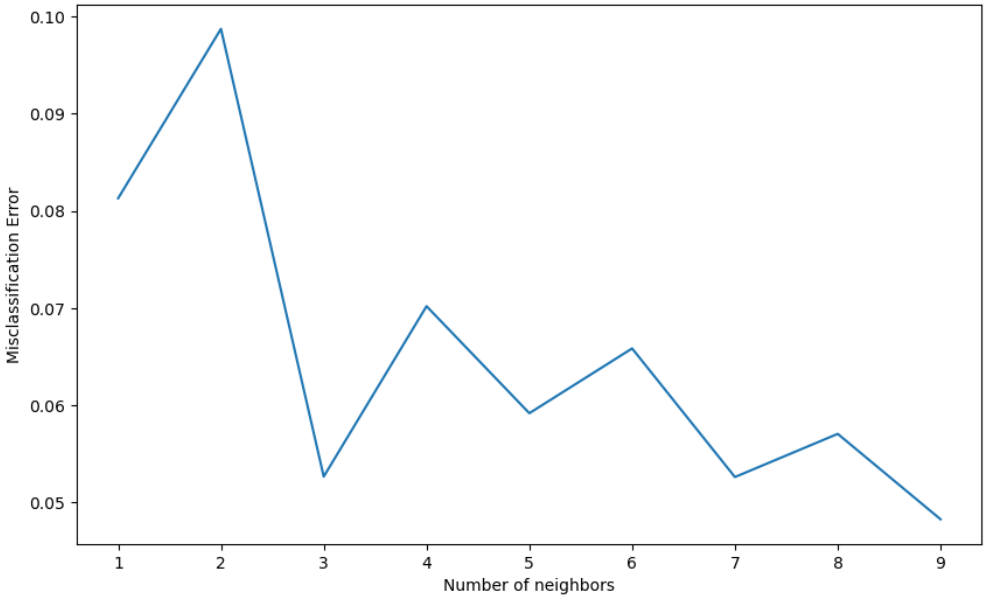


Based on the products of exploration, it was determined that all 10 features would be used in the training data. This was due in large part to the lack of one truly strong relationship between a feature and diagnosis.

To develop, train, and test the KNN model, SciKit-Learn was used. Data was broken into X- and y- testing and training groups. To ensure that all data was used in all stages, cross-fold validation was employed. To allow for continuous growth of the dataset, the cross-fold validation was carried through into the final model, meaning that the performance metrics of the model change slightly based on the given iteration.

As many of the features were on significantly different scales (for example, the mean value for smoothness is between 0.1029 and 0.09248 while the mean of area is 978.37 to 462.76), a scaler, also from the sklearn library, was used. In order to determine the optimal number of neighbors in the range of 1 to 10, a for loop was used, saving the cross-validation scores to a list. Using this, the miscalculation error (MSE) versus K was plotted and used to determine the optimal K value for the iteration. An example of this plot can be found at Figure II.

Figure II: MSE Given K Neighbors (Used for Optimization)



To view the performance of the model, a metrics classification report was used. This report shows numerous values, including precision, accuracy, f1-score, recall, and more. Preference was given to model accuracy due to the intended usage of the model. This is also the value that is reported to the user in the web application. An example classification report can be found at Figure III.

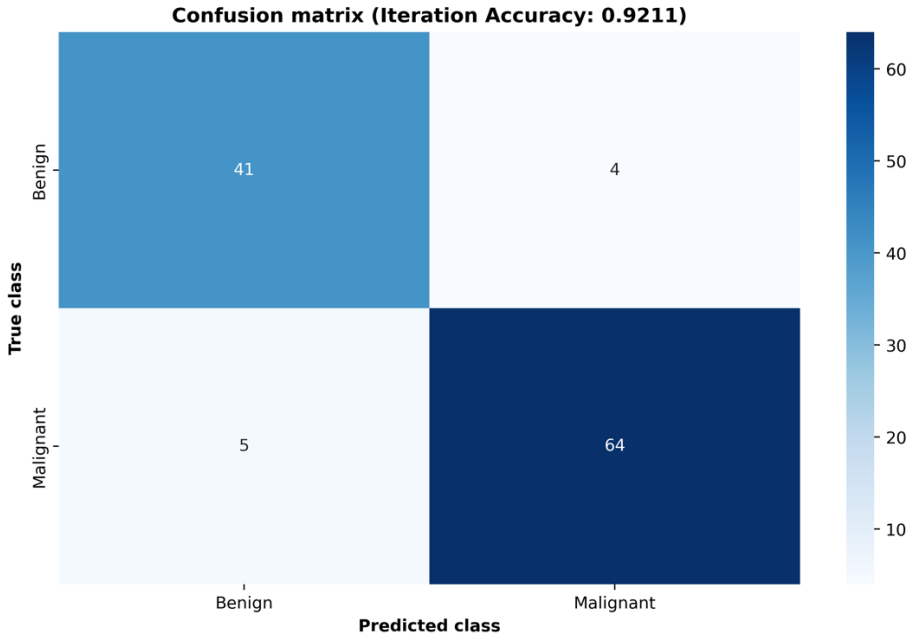
Figure III: Classification Report

	precision	recall	f1-score	support
0	0.89	0.91	0.90	45
1	0.94	0.93	0.93	69
accuracy			0.92	114
macro avg	0.92	0.92	0.92	114
weighted avg	0.92	0.92	0.92	114

accuracy is 0.9210526315789473

The final product made in this stage was a confusion matrix heatmap, which served as a representation of the correctly and incorrectly placed values. This was used to visually confirm that the model was working correctly. See Figure IV for an example of the plot.

Figure IV: Confusion Matrix Heatmap

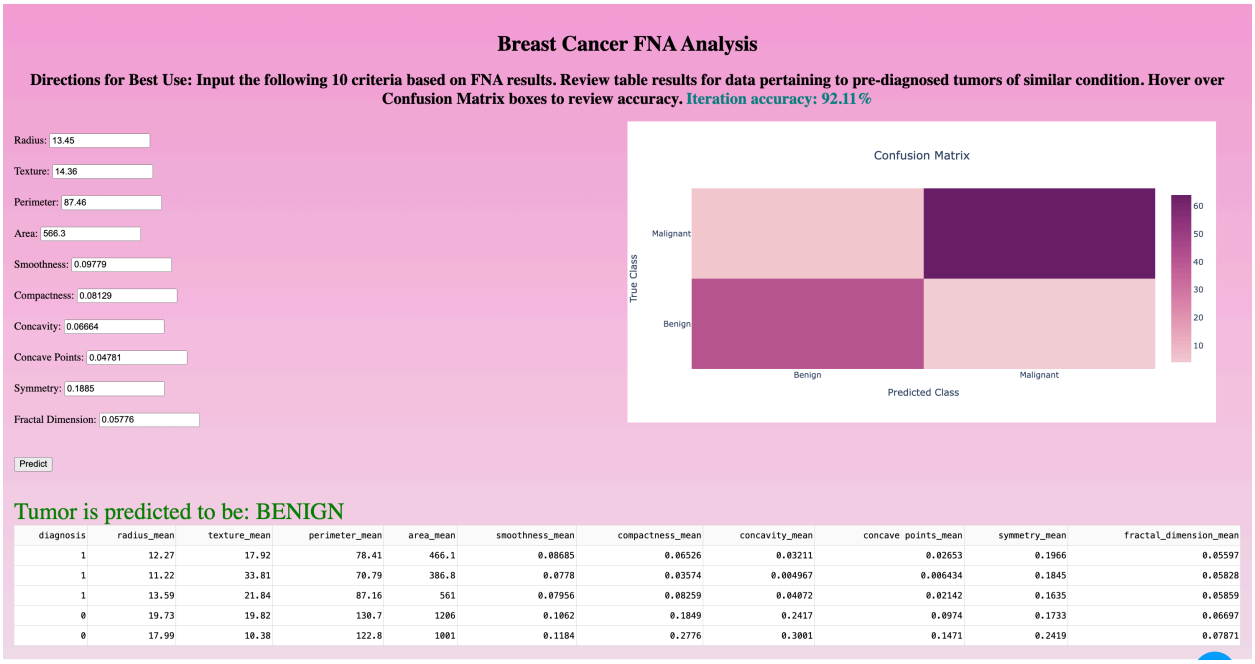


Conclusion and Final Product

In order to achieve the goal of creating a standalone platform to be used by non-technical persons, it was not appropriate to leave the user entry components as inputs in a Jupyter Notebook or similar file.

Initial consideration for the user interface was given to a Tkinter GUI or similar platform, creating a truly “stand alone” interface. However, due to design challenges, a Plotly Dashboard configuration was used. Value prompts are used to facilitate data inputs which then generate a prediction based on the KNN model. A confusion matrix heatmap (similar to the one shown in Figure IV) and table containing a sample of the nearest neighbors is returned as well. See Image I for a depiction of the final web application.

Image I: Final Web Application



Future Work

To extend on the work completed within the project, there are a number of further inclusions that could extend the reach and precision of the model and interface. One that I have considered is merging the Python aspects of the project with a backend database, created a pseudo- electronic medical records (EMR) system. This would both allow for enhanced record keeping on the providers' end and improve predictions made by the model via recycling of data (i.e. once the final diagnosis is entered, the features can be used to train new iterations).

Another possibility is to embed the dashboard and KNN model code within a cloud server, providing access from anywhere, with or without the code and an IDE installed. Again, this would further the reach and usage of the product, allowing providers to gain its insights without the need for exceptionally powerful technology.

Works Cited

1. https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html#:~:text=Breast%20cancer%20is%20the%20most,new%20female%20cancers%20each%20year.
2. https://seer.cancer.gov/statfacts/html/breast.html
3. Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., Shi, W., Jiang, J., Yao, P. P., & Zhu, H. P. (2017). Risk Factors and Preventions of Breast Cancer. International journal of biological sciences, 13(11), 1387–1397. https://doi.org/10.7150/ijbs.21635