Predicting the Risk of Breast Cancer

Capstone 3 Presentation by Joy Opsvig
April 2022

Introduction

The Problem

- Breast cancer has the second highest death rates among all cancers for women
- About 1 in 8 American women will develop invasive breast cancer over the course of her lifetime

Relevancy

- Early detection of malignant tumors is key to treating breast cancer patients
- If breast cancer is found early on, there are more treatment options available and higher chances of survival
- Women whose breast cancer is detected at an early stage have a 93 percent or higher survival rate in the first five years

Solution

- Use machine learning technology to accurately identify the diagnosis of breast cancer based on the measurements and attributes of a tumor
- Implement the model into a product available for medical centers, hospitals, and researchers to help patients identify early-stage breast cancer

Data Science Method

Approach for the predictive algorithm on malignant or benign tumor identification.

Steps:

1. Review and analyze features of benign vs. malignant tumors; do sanity checks.
2. Identify correlations between measurements and target variable (classification of tumor: benign vs. malignant).
3. Build predictive model on cleaned data, separating training and test data, after applied pre-processing techniques.
4. Perform model on test data and continue to improve model to achieve high accuracy rates with k-fold cross validation.
5. Ensure high accuracy so that inference on new data points identifies if tumors are at high risk of being classified as malignant with high recall.

Data Acquisition and Wrangling

The dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

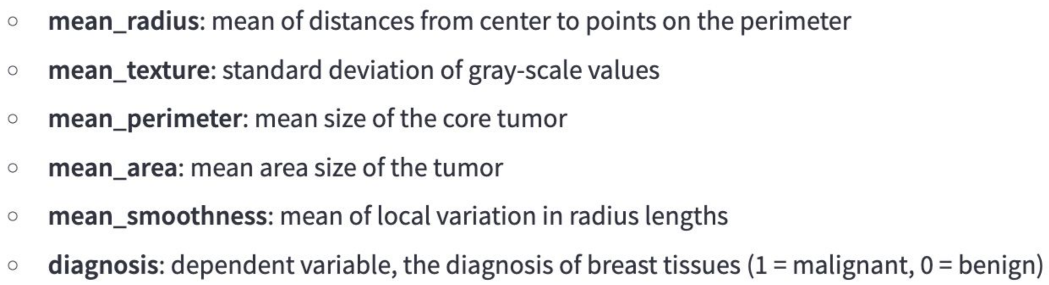
Data Summary:

- 569 entries of tumor measurements
- Five unique features
- Target variable: **diagnosis**

Data Overview:

- Data collected was mostly clean
- Outliers identified for the 'mean_smoothness' measurement; was removed

Data Dictionary:

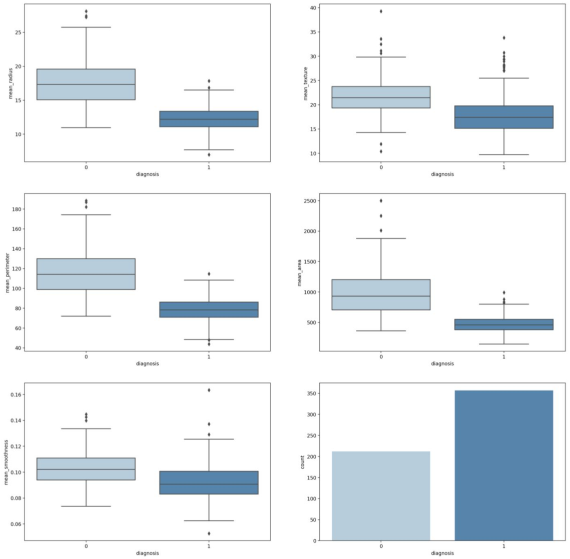


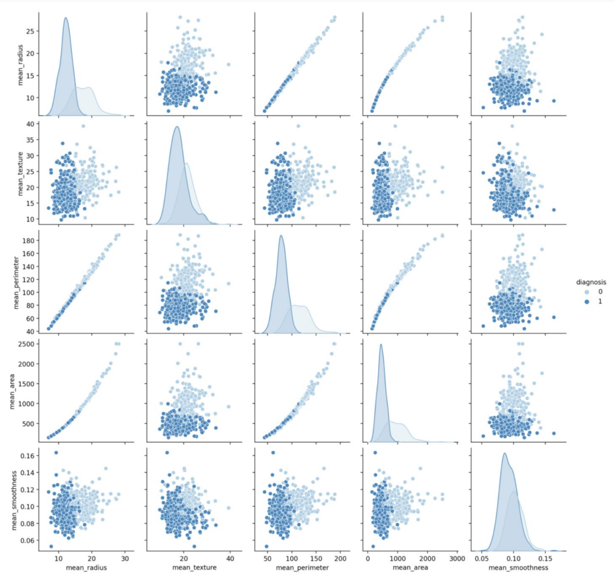
- **mean_radius**: mean of distances from center to points on the perimeter
- **mean_texture**: standard deviation of gray-scale values
- **mean_perimeter**: mean size of the core tumor
- **mean_area**: mean area size of the tumor
- **mean_smoothness**: mean of local variation in radius lengths
- **diagnosis**: dependent variable, the diagnosis of breast tissues (1 = malignant, 0 = benign)

Exploratory Data Analysis

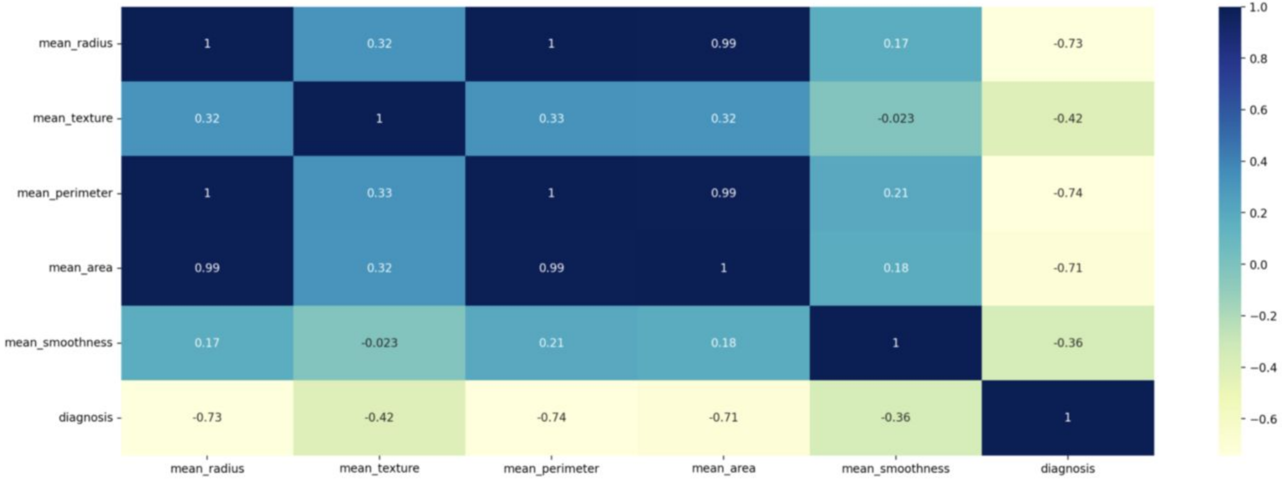
What we know:

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





Exploratory Data Analysis, con’t



Takeaways:

- As expected, 'mean_area', 'mean_radius', and 'mean_perimeter' are closely correlated with one another.

- Malignant tumors tend to have lower feature measurements, on average, compared to benign tumors.

Modeling

Built a supervised learning classification model to identify if a tumor is likely to be benign or malignant, and tested multiple modeling techniques & parameters to identify best performing.

Logistic Regression				accuracy: 89%
	precision	recall	f-1 score	support
0	0.97	0.70	0.81	50
1	0.86	0.99	0.92	92

Random Forest				accuracy: **94%**
	precision	recall	f-1 score	support
0	0.96	0.88	0.92	50
1	0.94	**0.98**	0.96	92

K-Nearest Neighbors				accuracy: 91%
	precision	recall	f-1 score	support
0	0.93	0.80	0.86	50
1	0.90	0.97	0.93	92

Model Considerations and Findings:

- Prioritize recall for metric evaluation in order to minimize false-negatives, i.e. incorrectly identifying a malignant tumor as benign.
- **Random Forest** scored highest for accuracy and higher across the board for the other metrics, including a high 98% recall for malignant tumors, only second to the Logistic Regression score.

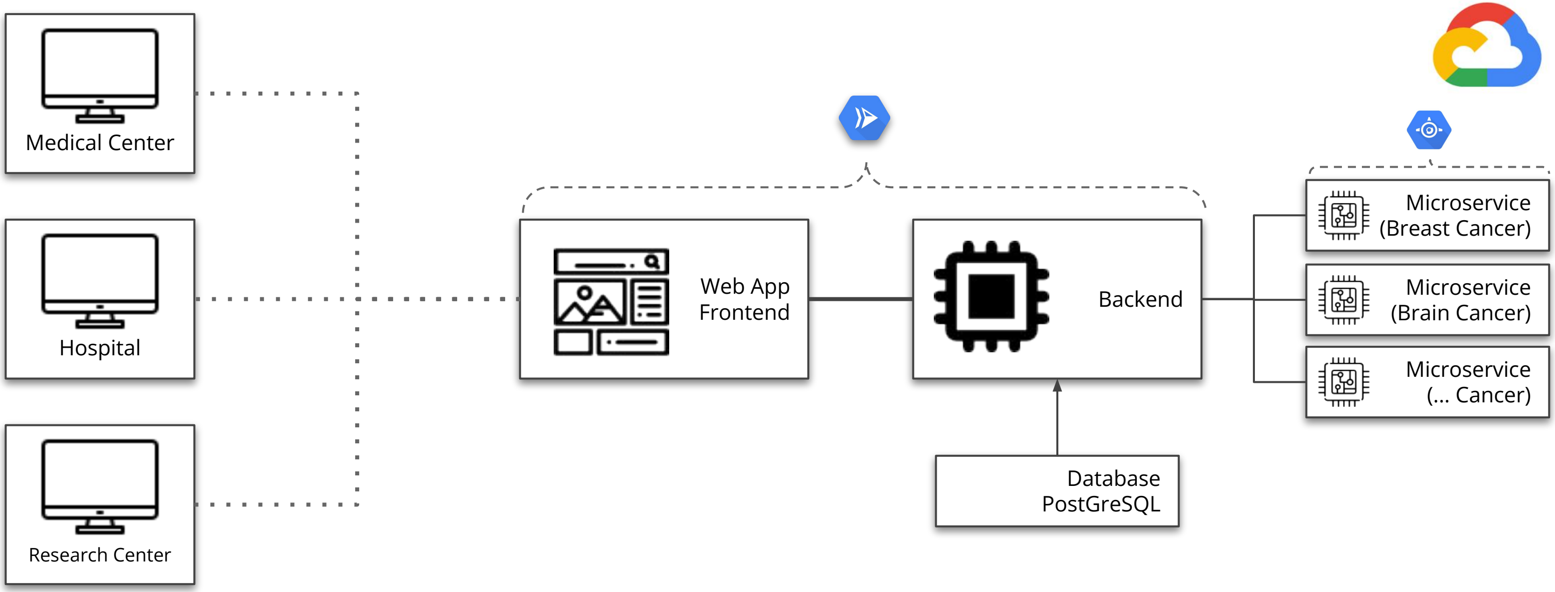
Extended Modeling

Additional steps for improving the model:

- Collect more data, upwards of 1,000,000 rows or more, as current dataset size is limited.
- Work with research team to identify additional features for measurement, retrain the model on these new features.
- Continue improving the model by implementing hyperparameter tuning via random search and grid search for the random forest model.

Integrating into IT landscape

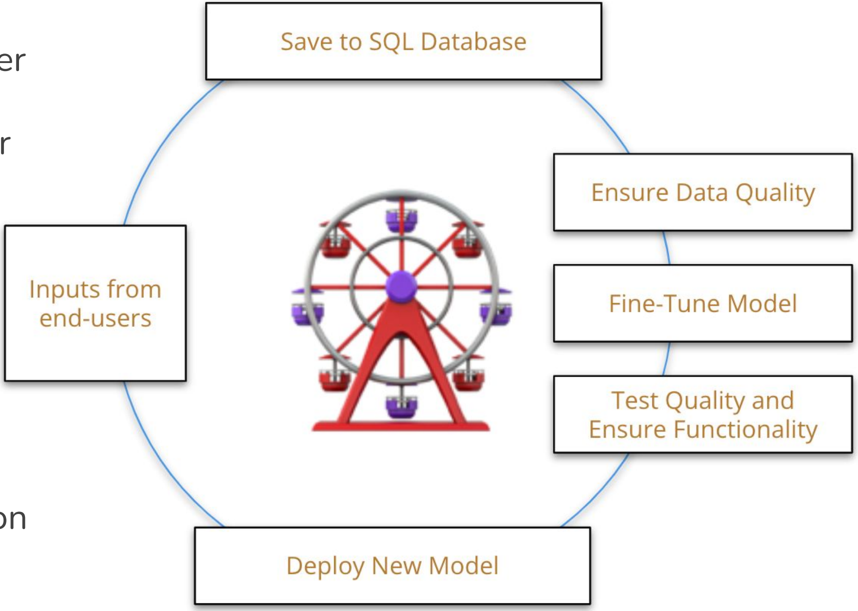
Allowing for interaction: Develop a web frontend for hospitals, medical centers, and research facilities to access trained models via API calls from backend to the microservices. Consider a database for credentials, and previous tumor data.



Medical Center
Hospital
Research Center
Web App Frontend
Backend
Database PostGreSQL
Microservice (Breast Cancer)
Microservice (Brain Cancer)
Microservice (... Cancer)

Integrating into Business Workflow

- **Commercialization:** Medical staff can enter patient data and determine whether a patient's tumor is likely to be malignant or benign. Charge on yearly subscription basis, monitor usage through API-call count.
- **Expanding product offer**: Develop and collect data points on different cancer types to allow for classifying additional tumor types.
- **Data flywheel:** Reuse data of users to ensure improvement of tumor identification models by continuous data ingestion (see right).



Conclusion

Client Recommendation

- Use with care: Particularly useful to indicate, by now doctors should not blindly trust.
- Extensive documentation on data quality standards and sources on website.
- Data privacy is ensured and abides by HIPPA.

Future Extensions of Product

- Support of additional cancer types (e.g., liver cancer).
- Analyze data for additional insights, such as identifying traits for people more likely to be at risk of cancer