Breast Cancer Diagnosis — Classification

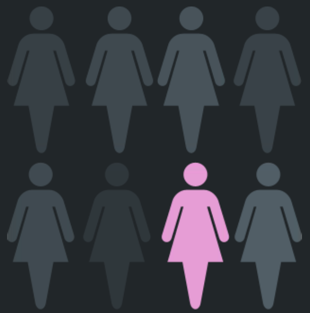
Ben van Zyll

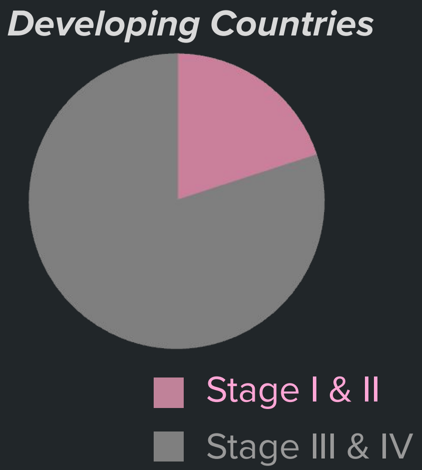
AGENDA

- Introduction & Overview
- Exploratory Data Analysis
- Baseline Model I — Logistic Regression
- Baseline Model II — Random Forest
- Baseline Model III — SVM
- Deep Learning — Feed-Forward Neural Network
- Conclusions & Future Experiments

OVERVIEW — The Problem

- **1 in 8** women will develop invasive breast cancer
- **30%** newly diagnosed cancer in women is breast cancer
- **70%** stage I and II diagnosis rate in developed countries
- **20%** stage I and II diagnosis rate in developing countries
- In 2022…
  - **287,850** invasive cases estimated
  - **43,250** estimated deaths





OVERVIEW — Our Approach

- **GOAL:** Create an optimal classification model
    - Tumor Diagnosis: Malignant / Benign

- Create three baseline models
    - Logistic Regression
    - Random Forest
    - SVM

- Create Deep Learning model
    - Feed-Forward Neural Network

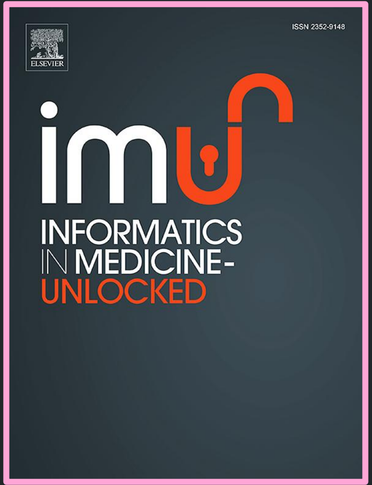
- Compare models, make conclusions

OVERVIEW — Current State of the Art Solution

- Common Diagnosis Procedures
    - Mammogram
    - MRI-guided biopsy
    - Ultrasound-guided biopsy

- *Informatics in Medicine Unlocked*
    - Logistic Regression Model
        - **95.71%** accuracy
        - **99.44%** sensitivity
        - **83.33%** specificity

- We seek to create a widely adaptable, cost-efficient method for tumor diagnosis



EXPLORATORY DATA ANALYSIS — Our Dataset

- Open-source, publicly available
- Shape: **569 x 21**

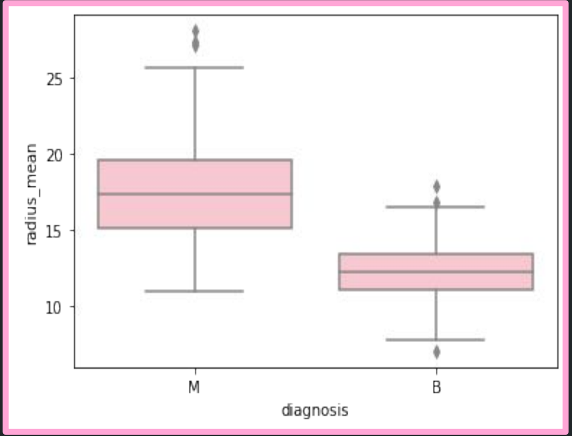
- Features:
  - radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst

- Label:
  - diagnosis

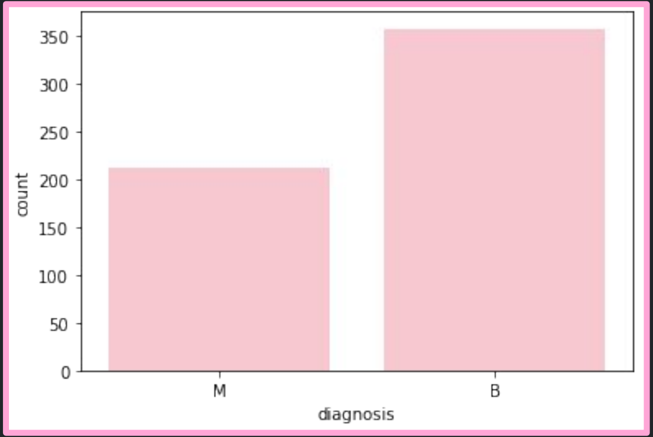
- No null values
- Encode binary outcome
  - Malignant : 1, Benign : 0

EXPLORATORY DATA ANALYSIS — Visualizations

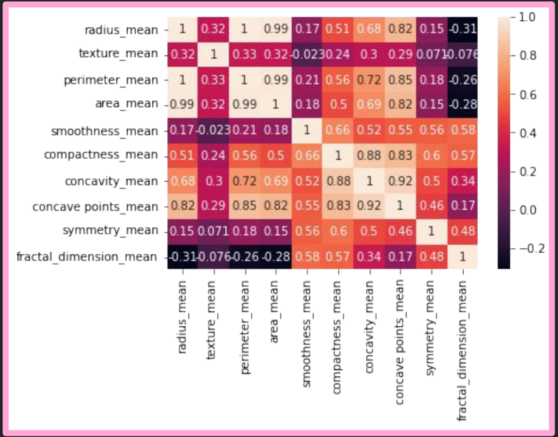
Boxplot



Diagnosis Barplot



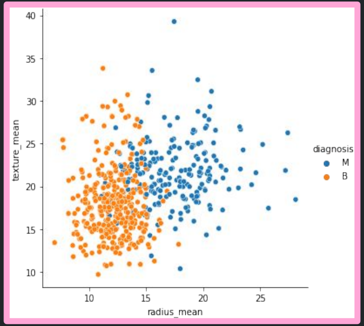
Correlation Matrix



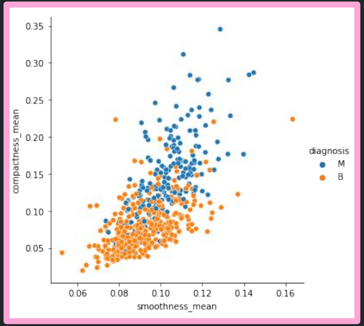
	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
radius_mean	1	0.32	1	0.99	0.17	0.51	0.68	0.82	0.15	-0.31
texture_mean	0.32	1	0.33	0.32	-0.023	0.24	0.3	0.29	0.071	-0.076
perimeter_mean	1	0.33	1	0.99	0.21	0.56	0.72	0.85	0.18	-0.26
area_mean	0.99	0.32	0.99	1	0.18	0.5	0.69	0.82	0.15	-0.28
smoothness_mean	0.17	-0.023	0.21	0.18	1	0.66	0.52	0.55	0.56	0.58
compactness_mean	0.51	0.24	0.56	0.5	0.66	1	0.88	0.83	0.6	0.57
concavity_mean	0.68	0.3	0.72	0.69	0.52	0.88	1	0.92	0.5	0.34
concave points_mean	0.82	0.29	0.85	0.82	0.55	0.83	0.92	1	0.46	0.17
symmetry_mean	0.15	0.071	0.18	0.15	0.56	0.6	0.5	0.46	1	0.48
fractal_dimension_mean	-0.31	-0.076	-0.26	-0.28	0.58	0.57	0.34	0.17	0.48	1

EXPLORATORY DATA ANALYSIS — Visualizations

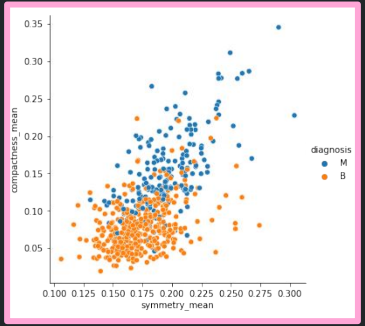
Radius / Texture



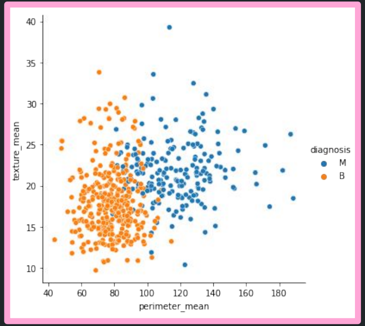
Smoothness / Compactness



Symmetry / Compactness

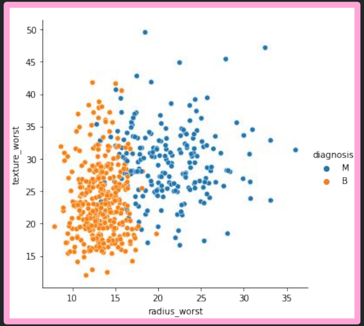


Perimeter / Texture

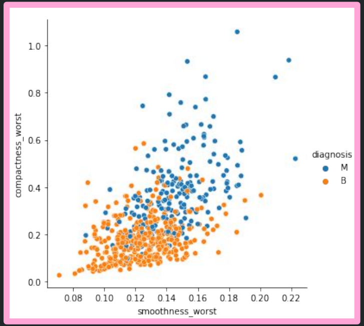


Means

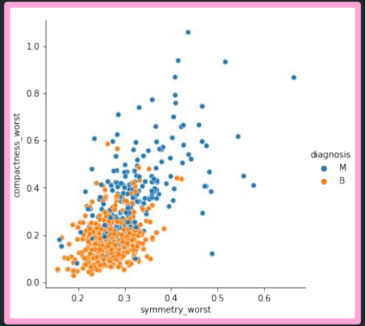
Radius / Texture



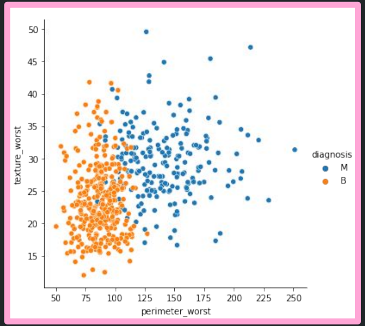
Smoothness / Compactness



Symmetry / Compactness



Perimeter / Texture



Worsts

BASELINE MODEL I — Logistic Regression

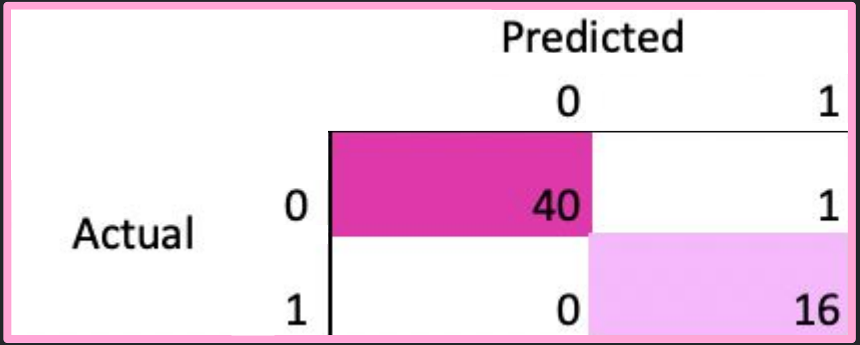
Training Accuracy: 0.952

Validation Accuracy: 0.982

Testing Accuracy: 0.982

Testing Precision: 0.941

Testing Recall: 1.0



		Predicted	
		0	1
Actual	0	40	1
	1	0	16

Optimal Hyperparameter: C = 10

BASELINE MODEL II — Random Forest

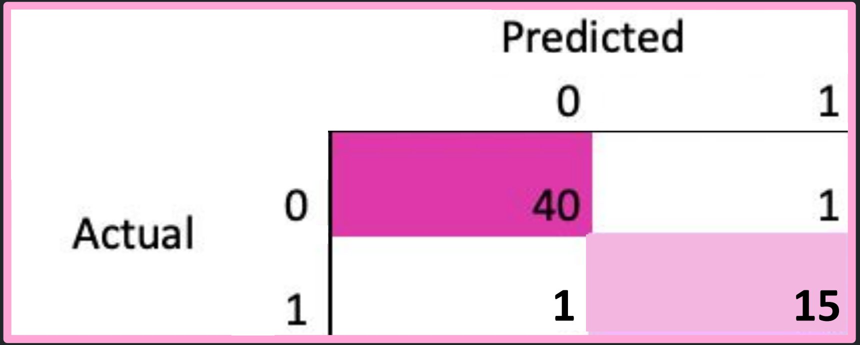
Training Accuracy: 0.965

Validation Accuracy: 0.965

Testing Accuracy: 0.965

Testing Precision: 0.938

Testing Recall: 0.938



		Predicted	
		0	1
Actual	0	40	1
	1	1	15

Optimal Hyperparameters: n_estimators = 100, max_depth = 8

BASELINE MODEL III — SVM

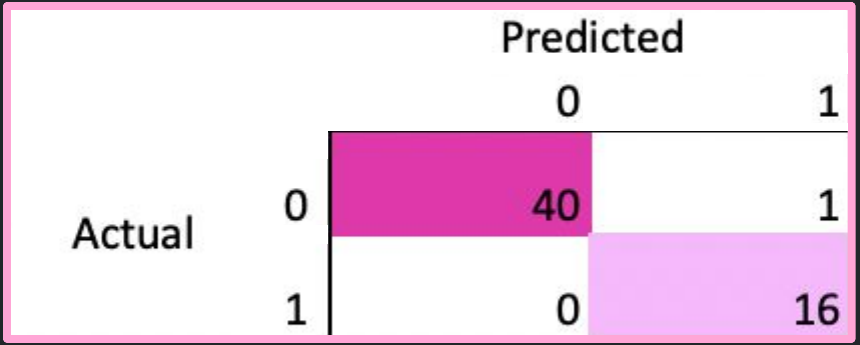
Training Accuracy: 0.95

Validation Accuracy: 0.974

Testing Accuracy: 0.982

Testing Precision: 0.941

Testing Recall: 1.0

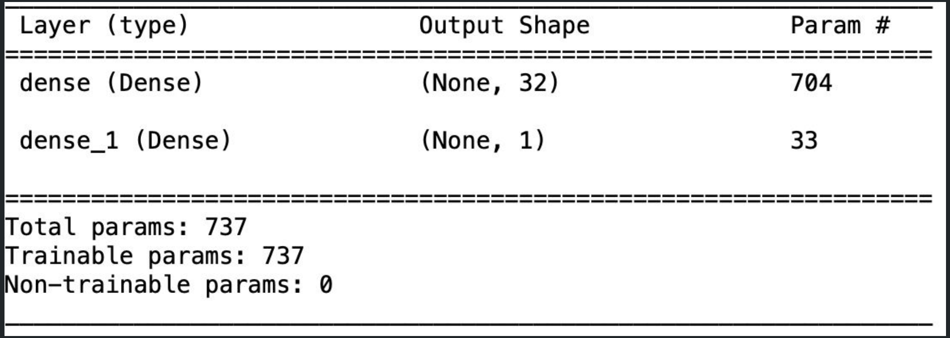


		Predicted	
		0	1
Actual	0	40	1
	1	0	16

Optimal Hyperparameters: C = 1, kernel = ‘linear’

Feed-Forward Neural Network — About the Model

- This is a sequential model that has 2 dense layers.
    - 1st layer - 32 units and uses the tanh activation function.
    - 2nd layer - 1 unit and uses the sigmoid activation function.
- Used Binary Cross Entropy as loss function
- Used L1 Regularization to prevent overfitting.

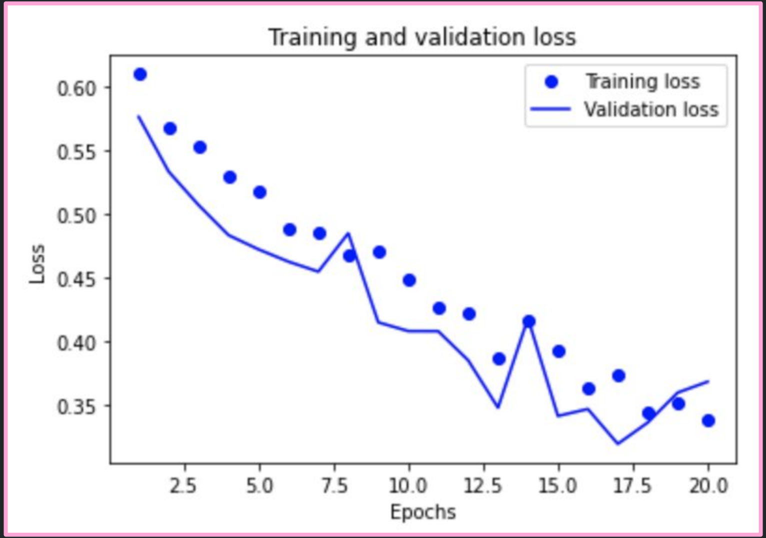


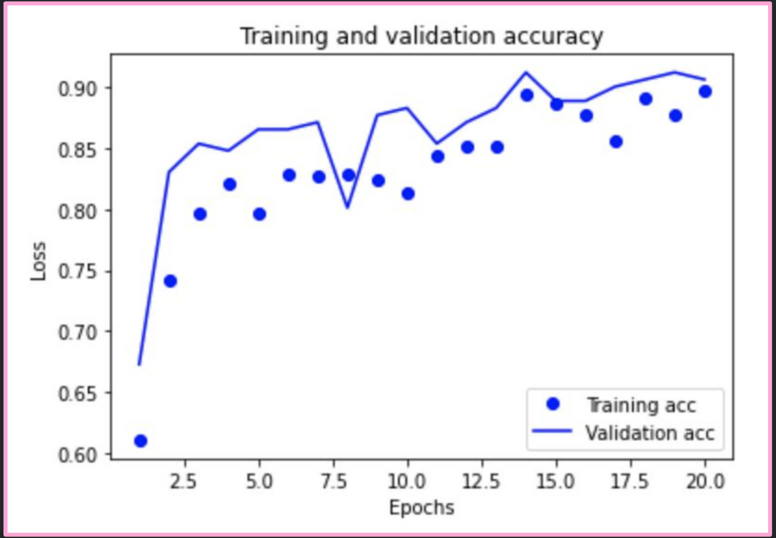
```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 32)                704

 dense_1 (Dense)             (None, 1)                 33

=================================================================
Total params: 737
Trainable params: 737
Non-trainable params: 0
_________________________________________________________________
```

Feed-Forward Neural Network — Results





Why might a Deep Learning Model perform worse?

- The neural network may be overfitting to the training data.
- The neural network may be poorly designed or configured.
- The baseline models are a better fit for the data
- The baseline models are less complex and therefore less likely to overfit
- The baseline model has fewer parameters and is therefore less likely to overfit

Future Experiments

- Need a cost-efficient, widely deployable technology that can provide measurements we see in the data
- Factor in other aspects of one's health/lifestyle in data
  - Age
  - Race
  - Income
  - Insurance
  - Diet
- Fine tune models to have lower-risk error
  - Lean toward false positive

References

- Challenges to the early diagnosis and treatment of breast cancer in developing countries. Karla Unger-Saldaña. World J Clin Oncol. 2014 Aug 10; 5(3): 465–477. Published online 2014 Aug 10. doi: 10.5306/wjco.v5.i3.465
- Breast Cancer Facts and Statistics. BreastCancer.org. 2022 March 10; Published online 2022 March 10. https://www.breastcancer.org/facts-statistics