Predicting Breast Cancer

A Comparison of Classification Models on Predicting Whether Breast Cancer is Malignant or Benign

by: Alyaqadhan Al Fahdi, Arya Amarnath

Data & Research Questions

Data

- The main source of data for this study will be the Breast Cancer Dataset, available on Kaggle (Breast Cancer Dataset, 2021)

- The Breast Cancer Wisconsin (Diagnostic) Dataset comprises of 570 entries with 32 columns. Each entry represents a case with features derived from a digitized image of a fine needle aspirate (FNA) of a breast tumor
    - radius_mean, perimeter_area, concavity_mean, etc

Research Questions

1. Can a historical dataset be used to classify the tumor type as 'M' indicating malignant (cancerous) or 'B' indicating benign (non-cancerous)?
2. Which model is most effective in classifying these tumors?

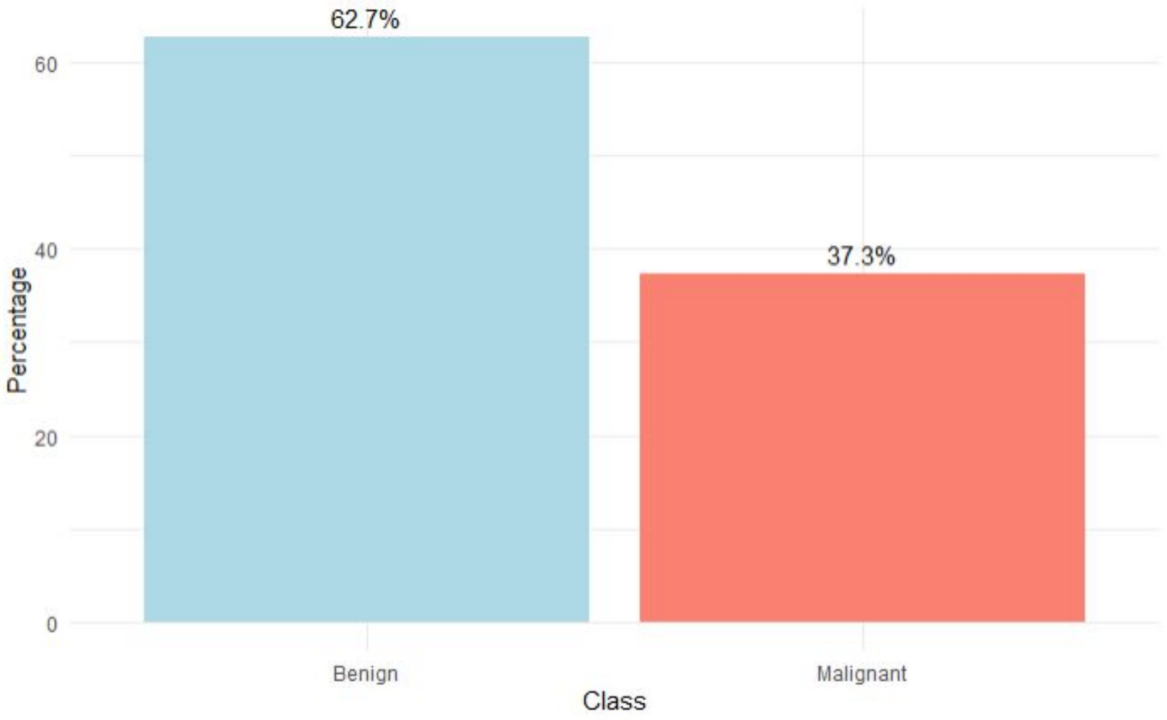
Methods

Prediction methods:

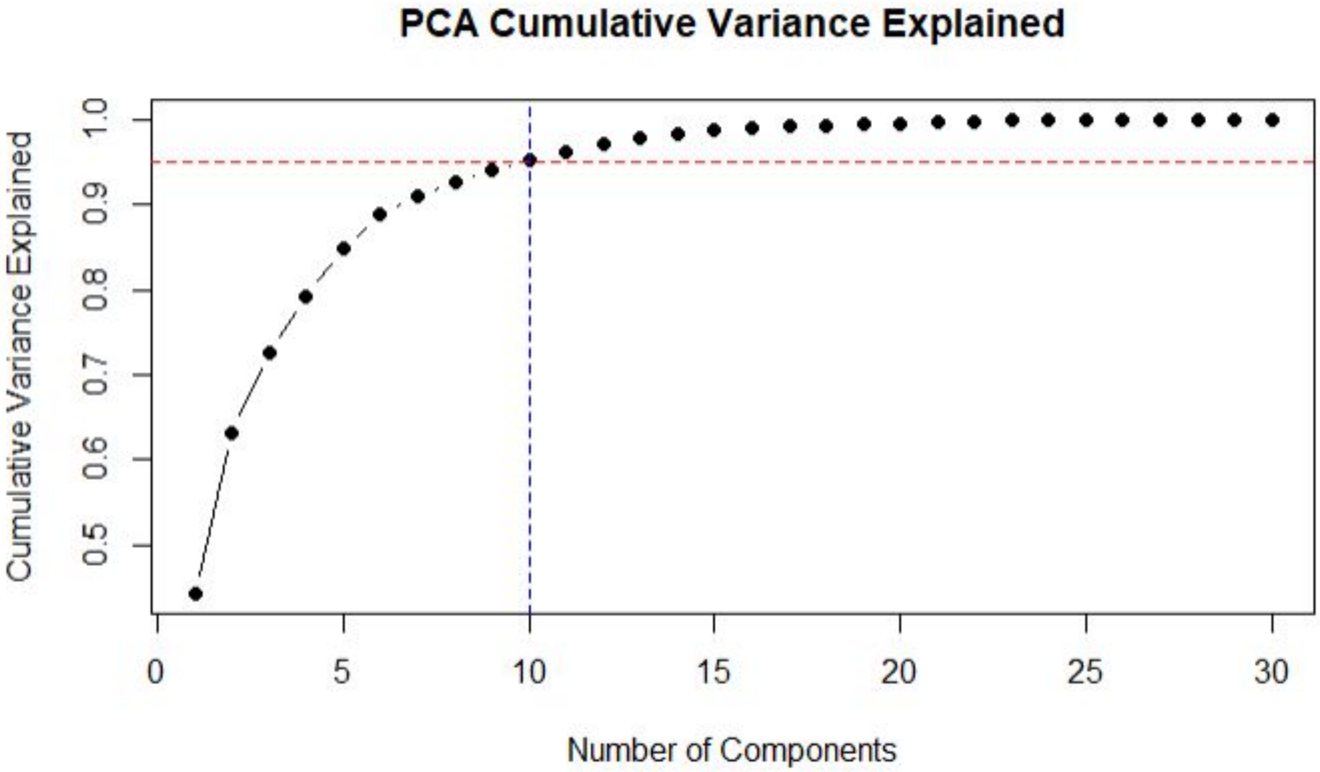
- Logistic
- SVM
- KNN
- Random Forest

Exploratory Data Analysis

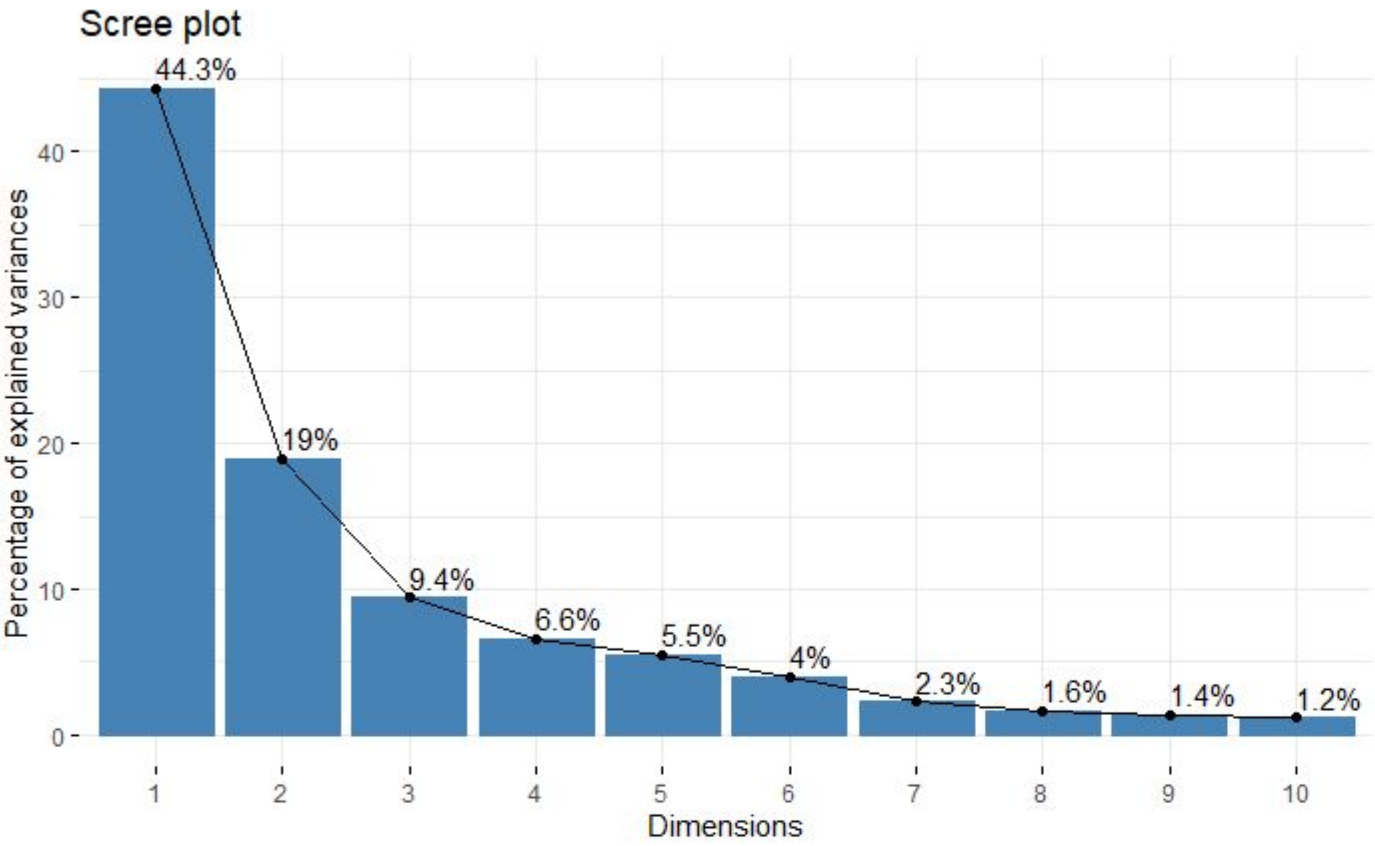
Exploratory Data Analysis

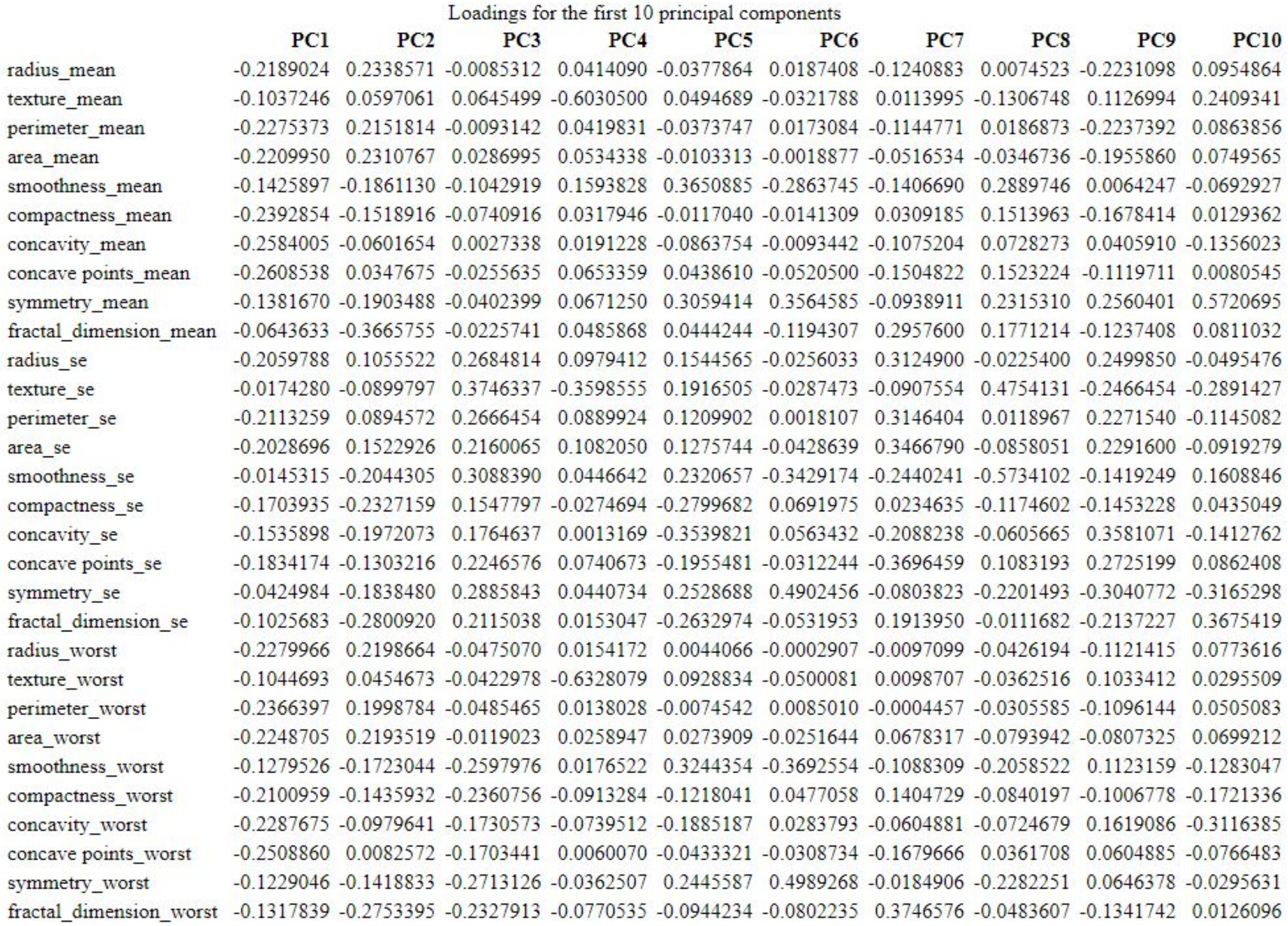


Principal Component Analysis (PCA)



PCA Cumulative Variance Explained


Scree plot
Percentage of explained variances
0
10
20
30
40
44.3%
19%
9.4%
6.6%
5.5%
4%
2.3%
1.6%
1.4%
1.2%
1
2
3
4
5
6
7
8
9
10
Dimensions



Loadings for the first 10 principal components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
radius_mean	-0.2189024	0.2338571	-0.0085312	0.0414090	-0.0377864	0.0187408	-0.1240883	0.0074523	-0.2231098	0.0954864
texture_mean	-0.1037246	0.0597061	0.0645499	-0.6030500	0.0494689	-0.0321788	0.0113995	-0.1306748	0.1126994	0.2409341
perimeter_mean	-0.2275373	0.2151814	-0.0093142	0.0419831	-0.0373747	0.0173084	-0.1144771	0.0186873	-0.2237392	0.0863856
area_mean	-0.2209950	0.2310767	0.0286995	0.0534338	-0.0103313	-0.0018877	-0.0516534	-0.0346736	-0.1955860	0.0749565
smoothness_mean	-0.1425897	-0.1861130	-0.1042919	0.1593828	0.3650885	-0.2863745	-0.1406690	0.2889746	0.0064247	-0.0692927
compactness_mean	-0.2392854	-0.1518916	-0.0740916	0.0317946	-0.0117040	-0.0141309	0.0309185	0.1513963	-0.1678414	0.0129362
concavity_mean	-0.2584005	-0.0601654	0.0027338	0.0191228	-0.0863754	-0.0093442	-0.1075204	0.0728273	0.0405910	-0.1356023
concave points_mean	-0.2608538	0.0347675	-0.0255635	0.0653359	0.0438610	-0.0520500	-0.1504822	0.1523224	-0.1119711	0.0080545
symmetry_mean	-0.1381670	-0.1903488	-0.0402399	0.0671250	0.3059414	0.3564585	-0.0938911	0.2315310	0.2560401	0.5720695
fractal_dimension_mean	-0.0643633	-0.3665755	-0.0225741	0.0485868	0.0444244	-0.1194307	0.2957600	0.1771214	-0.1237408	0.0811032
radius_se	-0.2059788	0.1055522	0.2684814	0.0979412	0.1544565	-0.0256033	0.3124900	-0.0225400	0.2499850	-0.0495476
texture_se	-0.0174280	-0.0899797	0.3746337	-0.3598555	0.1916505	-0.0287473	-0.0907554	0.4754131	-0.2466454	-0.2891427
perimeter_se	-0.2113259	0.0894572	0.2666454	0.0889924	0.1209902	0.0018107	0.3146404	0.0118967	0.2271540	-0.1145082
area_se	-0.2028696	0.1522926	0.2160065	0.1082050	0.1275744	-0.0428639	0.3466790	-0.0858051	0.2291600	-0.0919279
smoothness_se	-0.0145315	-0.2044305	0.3088390	0.0446642	0.2320657	-0.3429174	-0.2440241	-0.5734102	-0.1419249	0.1608846
compactness_se	-0.1703935	-0.2327159	0.1547797	-0.0274694	-0.2799682	0.0691975	0.0234635	-0.1174602	-0.1453228	0.0435049
concavity_se	-0.1535898	-0.1972073	0.1764637	0.0013169	-0.3539821	0.0563432	-0.2088238	-0.0605665	0.3581071	-0.1412762
concave points_se	-0.1834174	-0.1303216	0.2246576	0.0740673	-0.1955481	-0.0312244	-0.3696459	0.1083193	0.2725199	0.0862408
symmetry_se	-0.0424984	-0.1838480	0.2885843	0.0440734	0.2528688	0.4902456	-0.0803823	-0.2201493	-0.3040772	-0.3165298
fractal_dimension_se	-0.1025683	-0.2800920	0.2115038	0.0153047	-0.2632974	-0.0531953	0.1913950	-0.0111682	-0.2137227	0.3675419
radius_worst	-0.2279966	0.2198664	-0.0475070	0.0154172	0.0044066	-0.0002907	-0.0097099	-0.0426194	-0.1121415	0.0773616
texture_worst	-0.1044693	0.0454673	-0.0422978	-0.6328079	0.0928834	-0.0500081	0.0098707	-0.0362516	0.1033412	0.0295509
perimeter_worst	-0.2366397	0.1998784	-0.0485465	0.0138028	-0.0074542	0.0085010	-0.0004457	-0.0305585	-0.1096144	0.0505083
area_worst	-0.2248705	0.2193519	-0.0119023	0.0258947	0.0273909	-0.0251644	0.0678317	-0.0793942	-0.0807325	0.0699212
smoothness_worst	-0.1279526	-0.1723044	-0.2597976	0.0176522	0.3244354	-0.3692554	-0.1088309	-0.2058522	0.1123159	-0.1283047
compactness_worst	-0.2100959	-0.1435932	-0.2360756	-0.0913284	-0.1218041	0.0477058	0.1404729	-0.0840197	-0.1006778	-0.1721336
concavity_worst	-0.2287675	-0.0979641	-0.1730573	-0.0739512	-0.1885187	0.0283793	-0.0604881	-0.0724679	0.1619086	-0.3116385
concave points_worst	-0.2508860	0.0082572	-0.1703441	0.0060070	-0.0433321	-0.0308734	-0.1679666	0.0361708	0.0604885	-0.0766483
symmetry_worst	-0.1229046	-0.1418833	-0.2713126	-0.0362507	0.2445587	0.4989268	-0.0184906	-0.2282251	0.0646378	-0.0295631
fractal_dimension_worst	-0.1317839	-0.2753395	-0.2327913	-0.0770535	-0.0944234	-0.0802235	0.3746576	-0.0483607	-0.1341742	0.0126096

Model Evaluation

Model Setup

Data Split:

- Training set: 70%
- Testing set: 30%

Hyperparameter Tuning:

- Cross-Validation (7 Fold)
- Grid Search

Evaluation Metrics:

- Balanced Accuracy
- Sensitivity
- Type II Error

Evaluation Metrics

Why **Balanced Accuracy, Sensitivity, Type II Error?**

Balanced Accuracy provides a more honest evaluation when there is an imbalanced dataset.

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

Sensitivity measures the proportion of actual positives (malignant cancers) that are correctly identified.

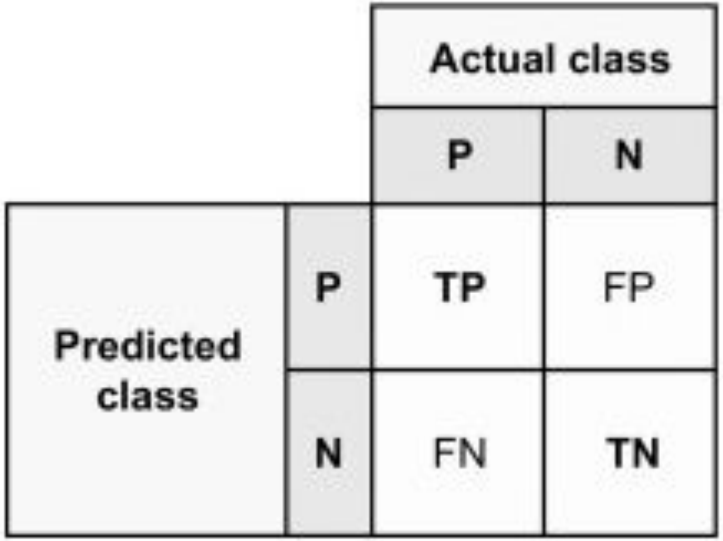
$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Type II error, also known as the False Negative Rate, is the proportion of positives (malignant tumors) that produce negative test outcomes (incorrectly identified as benign).

$$\text{Type II Error} = \frac{FN}{TP+FN} = 1 - \text{Sensitivity}$$

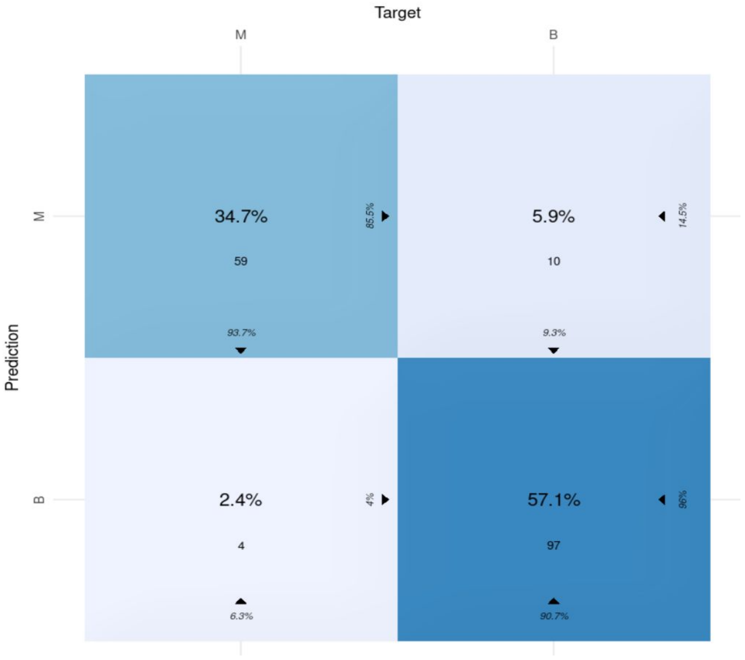
Confusion Matrix

- ***TP*** **(True Positives)** is the count of malignant tumors correctly identified.
- ***TN*** **(True Negatives)** is the count of benign tumors correctly identified.
- ***FP*** **(False Positives)** is the count of benign tumors incorrectly identified as malignant.
- ***FN*** **(False Negatives)** is the count of malignant tumors incorrectly identified as benign.

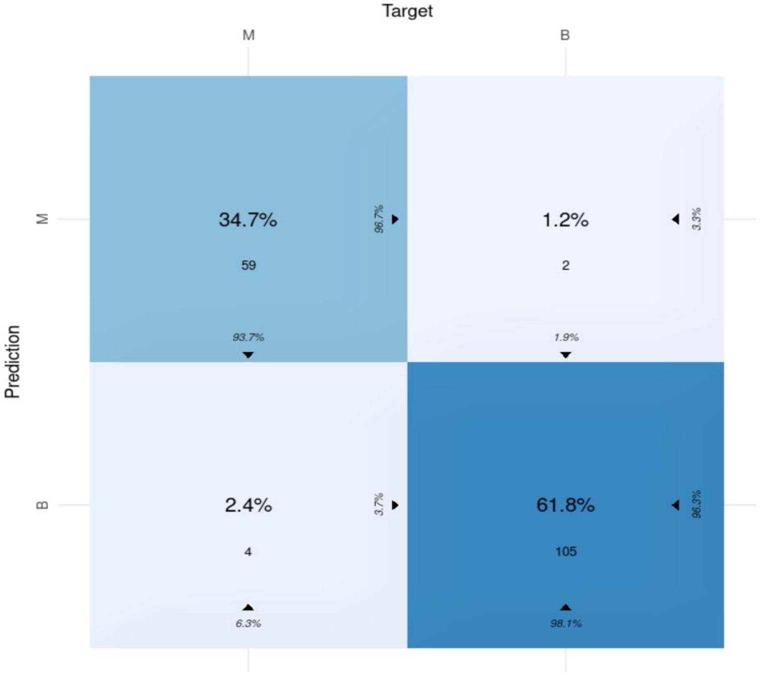


		Actual class	
		P	N
Predicted class	P	TP	FP
	N	FN	TN

Evaluation

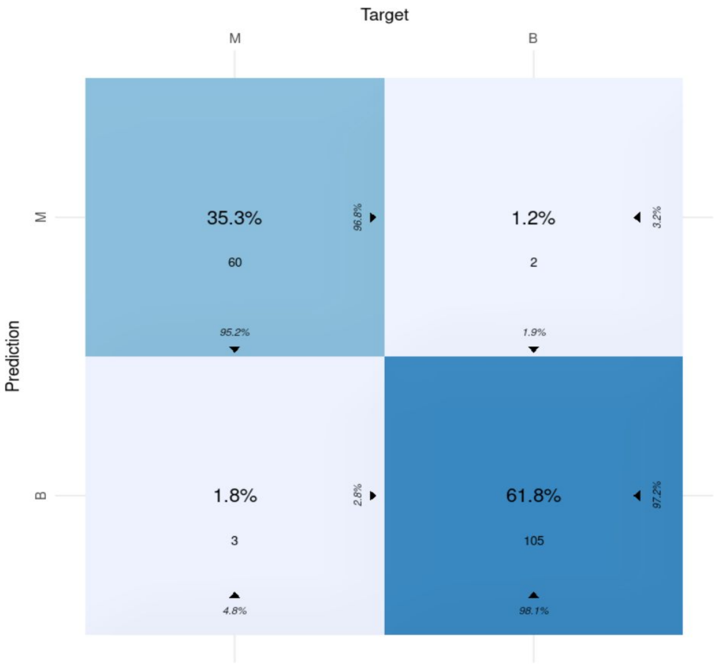


Random Forest

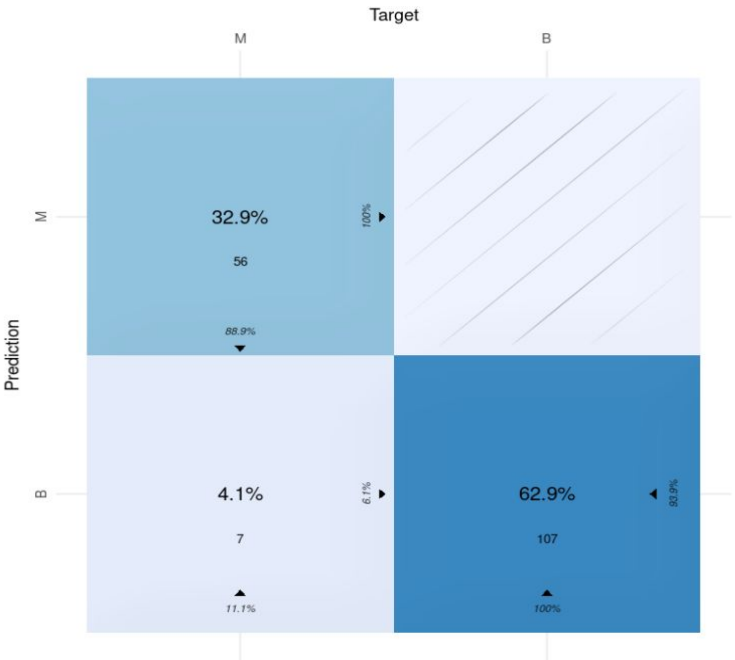


SVM

Evaluation

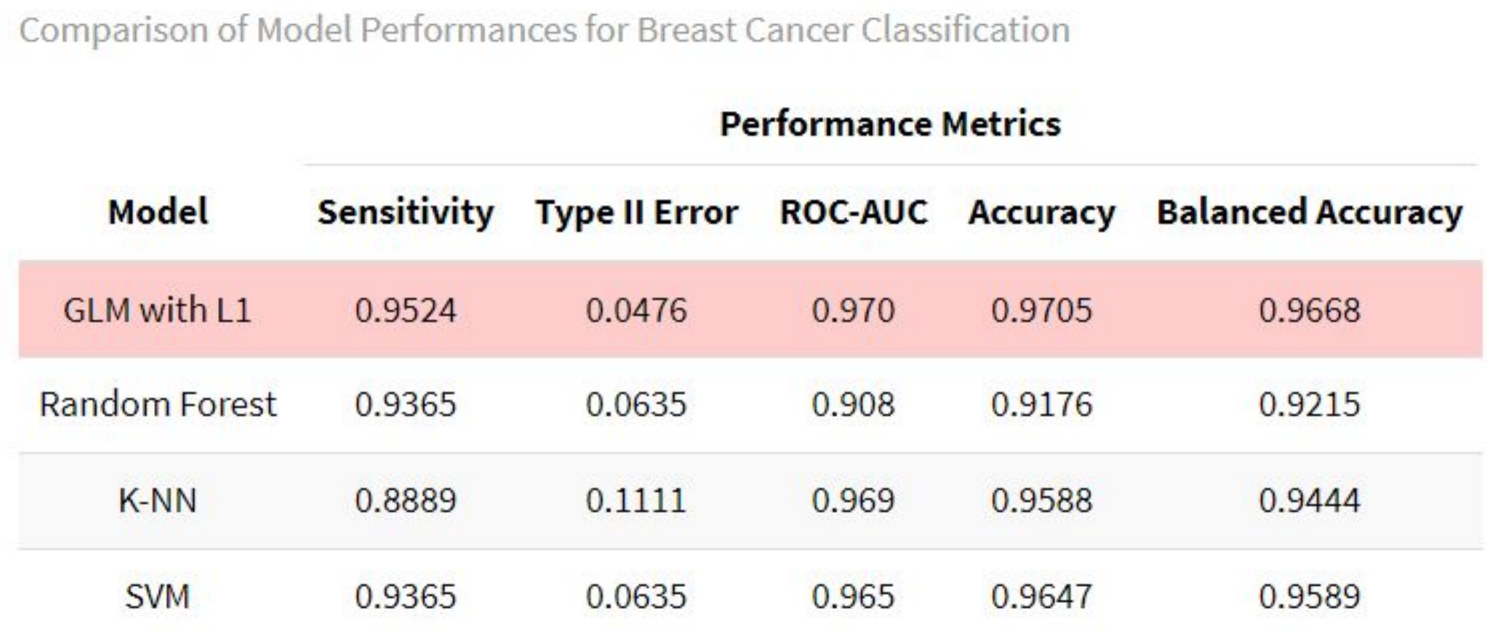


Logistic



K-NN

Model Comparison



Comparison of Model Performances for Breast Cancer Classification

	Performance Metrics				
Model	**Sensitivity**	**Type II Error**	**ROC-AUC**	**Accuracy**	**Balanced Accuracy**
GLM with L1	0.9524	0.0476	0.970	0.9705	0.9668
Random Forest	0.9365	0.0635	0.908	0.9176	0.9215
K-NN	0.8889	0.1111	0.969	0.9588	0.9444
SVM	0.9365	0.0635	0.965	0.9647	0.9589

Conclusion

Historical Data as a Foundation: Utilizing historical data from breast cancer cases provides a valuable perspective for predicting the nature of the disease, malignant or benign

PCA for Complexity: Principal Component Analysis (PCA) served as an effective tool for reducing the complexity, reducing feature redundancy while maintaining 95% of the original variance within 10 principal components

General Logistic Model w/ L1 Regularization: The GLM with L1 regularization became the premier model in our study. It showed an excellent performance with a **Sensitivity of 95.24%**, **Type II Error Rate at 4.76%**, and an overall **Balanced Accuracy of 96.68%**, indicating its accuracy in classifying the cancer types almost flawlessly

Future Horizons in Cancer Prediction: Building on the GLM with L1 model's success, future explorations could involve cross-validation with larger datasets. Possible factors such as adding genetic and lifestyle factors would help increase predictive accuracy and provide a more specific risk analysis

References

- *Breast Cancer Dataset*. (2021, December 29). Kaggle. Breast Cancer Dataset
- Mitrani, A. (2021, December 12). Evaluating Categorical Models II: Sensitivity and Specificity. *Medium*.

  https://towardsdatascience.com/evaluating-categorical-models-ii-sensitivity-and-specificity-e181e573cff8