BREAST CANCER CLASSIFICATION WITH MACHINE LEARNING ALGORITHMS

TABLE OF CONTENTS

- About Project
- Goals
- Dataset Overview
- Exploration Data Analysis
- Models Selection
- Result
- Best Model Analysis
- Conclusion



Naftalia Sophiana Purba

Data Scientist | Data Analyst | Machine Learning Enthusiast

P Bali, Indonesia

Undergraduate student of Mathematics from Udayana University with a deep interest in Data Analytics, Data Science, and Machine Learning

TOOLS AND LIBRARY













ABOUT PROJECT

This project aims to build a machine learning model to classify breast tumors as benign or malignant using the scikit-learn dataset. The tested models include SVM, Logistic Regression, Random Forest, KNN, and Naïve Bayes, compared to determine the best performance in breast cancer classification.

GOALS

The primary objective of this project is to develop an accurate and efficient machine learning model for breast cancer classification using the Scikit-learn dataset. By applying data preprocessing, feature analysis, and evaluating multiple classification algorithms, the project aims to enhance predictive performance and support early cancer diagnosis to assist medical professionals in making informed decisions.

DATASET OVERVIEW

DATASET OVERVIEW

• Source : Scikit-learn

• Number of Instances : 569

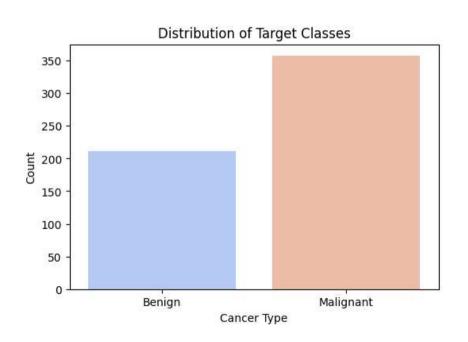
• Number of Features : 30

Feature Types : Numerical

• Traget Classes : 0 = Benign (No cancerous tumor)

1 = Malignant (Cancerous tumor)

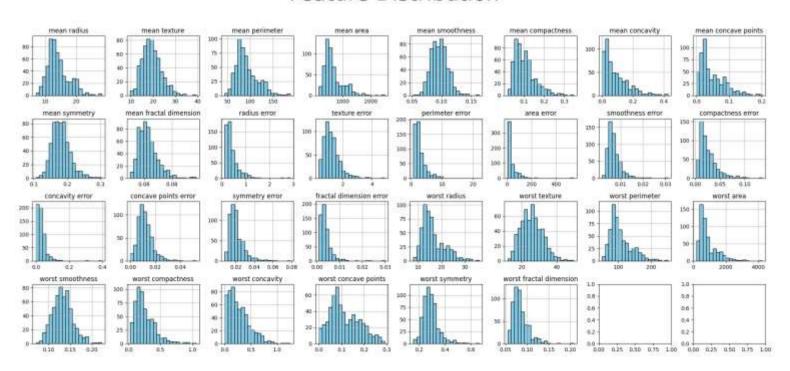
EXPLORATION DATA ANALYSIS



The class distribution in the graph shows that the number of malignant (cancerous) cases is higher than benign (non-cancerous) cases. This provides insight that breast cancer is more frequently detected as malignant in this dataset.

EXPLORATION DATA ANALYSIS

Feature Distribution



EXPLORATION DATA ANALYSIS

								_	Cor	WINNESS	The same of	NO SHARE		delicated in		Appellation		SHOW	Indian's	Delica Control	HERMAN	Section 200	-	-								
THE PERMIT	3.00	835	111	77	0.21	ZM	1100	100	90,000	200	100	0.00	Lang.	555	-34	833	1110	10.00	H-14	377	0.007	100	100	100	SEEK.	MATE	444	100	110	6363		
mean testure	nu.	1.01	1111	9.86	411	894	10-00	120	201	0.00	9.28	1.20	nie i	6.21	0.01	619	1114	0.46	0.01	4.84	4.00	***	d.m	5.14	430	219	0.00	1129	1000	6.11	-416	
man partitions	2.00	0.44	100	***	0111	2.34			75.65	0.00		0.00	-		0.04	9.22	200	-	414	-0.01	775	9.90	2.77		10.631	MAN	871	111	4.00	40.00		
FR81 2700	0.00	0.21	0.00		0.10	200	-	180	Kith	4.00	2011	0.01	100	***	-634	8.01	16.24	0.00	4.00	46.00	496	020	-	***	MEE	tide:	9.91	211	THE R.	****		
oge productivation	11.11	0.000	mas	***	1.00	100	ent	399	0.00			637	-	633	400	8.11	220	1,000	12.000	(998)	0.31	1886	0.34	331	444	MAT	0.60	525	4.94	4111	***	
e conquirtues	0.01	0940	+034	0.00		4.84	-		-	***	a.lm	6.05	Miles	0.00	4.40	***	801		0.21	4.81	nie	936		430	400	***	***				-	
main company.		0.00	MIN)	***	000	***	-	777	-100	534	-	100			20.60	***			0.00	***		996	200	1000	4.00	***	177	***		981	277	
cocesse power	0.00	0.24	1000	-	-	***	100	-189	7.00	8.00	100	890	WHI.	(44)	0.00	1.00	2:00	***	0.00	0.04	481	929	-	***	200		411	491	10.00	0.24	0.10	
name systematry	035	1000	9306	277	19.04		1100	2.17	1.00	***	-	9-53	1000	944	20100	221	1004	0.94	0.00	(834)	410	Nee.	0.50	4.19	19.00	11.11	0.45	1149	857	0.04	677	
actel directors	#.31	0.00	414	m.14		0.01	934	*11	1000	100	0.000	***	0.04	40.00	0.40	4.11	0-47	0.34	420	***	+21	A100	0.21	4.01	430	1140	0.30	0.10		440	0.01	
talke are		0.00	-	386	0.00	0.04	9	155	4,00	***	100	P-05	0.00	200	9040	4.00	0.00	***	1915	834	486	030		32	4,64	+4+	0.00	***	277	41.65	199	
SECRETARISM STORY	- 31	0.11	***	-000	9101	0.01	19448	490	9.69	***	9-94	180	30.00	***	0.00	0.11	414	0.03	0.01	104	440	100	4-14	30.04	-47	4.00	0.47	34.60	4.17	41.91	444	
(81) 1134 117 117 117	100	0.34			010	+01			-0.00	***	0.00	+50	100		948	***	N:30	0.04	0.(1)	1885		0.00	Sept.		410	0.04	041	244	-11	0.04	1705	
-	Hile	0.10	414	***	400	0.44	0 40	244	*11	4.00	4	3111	***	110	0.00	200	4.01	0.42	0.00	*10	144	830	38	440	4.60	*11	0.94	134	441	0.02	G-34	
CONTRACTOR	441	1006	0.00	-	0.00	0.04	0.00	430		140	910	5.60	0.00	911	+1=	411	441	0.94	0.003	9.41	4.03	1000	6.00	***	**	0.00	0.04	11.10	1915	646	0.01	
speciment stee	2124	-11	1100	*36	0.00	*		25.40	9.44	444		649	9.46	611	10.00	+			0.00	***	630	0.04	1934	430	4.61	***		0.00	-0.10		434	
emessing etter	038	0.94	1669.	200	1000	+91		2.40	-	***	0.54	839	0.40	0.01	4.01		100	+77	0-01	***	444	030	933	E49.	9.65	1548	-	***	-10	0.04	0.00	
	li in		1000	+41	4110	yyn	-	100		2.24	441	8.00	-	6.43	3.50	-10		188	2.25		5.80	540	10.00	X10.	420	646	446		19.64	66	-0-14	
	330	896	0.00	-0.07	000	5-31	0.46	830	0.00	4.46	034	0.44	277	844	0.01	0.00	3000	034	1-86	9.41	0.19	-0.00	9.18	22.55	9.01	0.00	0.00	0.01	-11	614	.000	
******	444	800	0.01	4.46	nin	344	000	838	-116	***	636	120	Air.	923	046	***	100	(913)	4.07	+44	0.00	-	490	30.60	441	4.44	600	449	-11		0.00	
ment orbat	5.61	0.00	-		en.	0.04	9	***	mer	8.00	10.000	11.11		100	-6-26	0.00	110	0.94	200	-0.04	-	10.001	-	***	100	4.40	0.01	***	-0.00	***	-7-	
Mark the barrier	16.00	***	0.00	444	-0.00	8-24	930	228	0.00	4.00	0614	11.00	900	6.81	3640	***	440	0.99	4.66	-	6.40	1.60	0.27	20.41	6.01	3-4m	9.00	11310	-0.04	0,00		
more of particular	0.01	0.74	0.01	***	614	100	2.72	***	0.41	0.01	10.71	H 10	-		-6-20	9-05	777	0.94	0.18	-0.00	***	0.07	-	***	835	440		***		644	0.70	
septial name	-	-	0.00		931	A.44	-	100	10,140	0.00		0108	400	146	0.38	0.22	410	0.10	4.60	8.00	-	ARE	-	110	0.00	4.44	9.94	HB	(64E)	0.04	-10	
and sensellarane	340	1995	8.80	944	inn.	1000	0.00	.040	0.00	4.65	0.14	0.01	810	5.11	-	5.81	447	0.55	0.00	945	9.83	831	0.24	349)	1.00	447	0.00	200	-		241	
	3,40	ADE.	11-10	0.00	0.00	***	-11		0.47	844	6.54	***	ALM:	0.22	0.00		-	0.40	0.00	0.00	0.00	A-10	-9.00	0.40	4.67	110					240	
world nonempty	0.00	0.00	1110		0.45	1.01	100	3.75	0.00	0.00	0.00	0.01	0.00	0.01	(0.00)	-	1		0.00	0.00	***	2.47		0.04	4.55		100		200		***	
COCCAVE JUSTIN	(412)	A.94	446		1000	***	-	100	18.61	226	440	nus	Arm.	0.04	*31	-0.00	0-40		4.60	0.22		2.00	100		970		-	110		4.00	A 19	
ment sprometry	1110	876	9.80	711	12.00	10.01	0.46	710	-	8.00	-	11.10	814	0.01	431		-	0.86	0.00	9.66	0.16	520	0.07	3.80	10.00		441	100	1-	20.00	317	
NAME OF STREET	0.00	AZZ.	0000	446	410	100	HIL	811	1000	4.00	400	0.00	ALON	6.00	30.60	10.0	200	6.00	0.06	1111	10.00	1999	934	0.34		100	(C##)	223	-10.00	100	A.44	
94540	4.11	11/10	2.00	-71	41.10	2	2.77	100	-0.00	100	3317	2.05	100	553	881	200	9.41		10.01	20.00	4.00		4.11	10.73			1000	0.79		9.17	1.00	
	- 16		100	4	140	1.00		- 40	200	. 4	1.6	16.7	1	160	- 64	90.0	146	Ber C	10.		1	11 e C -				- 0	100	4		-		



MODEL SELECTION

- Support Vector Machine (SVM)
- Logistic Regression
- Random Forest
- K-Nearesr Neighbour
- Naive Bayes

RESULT

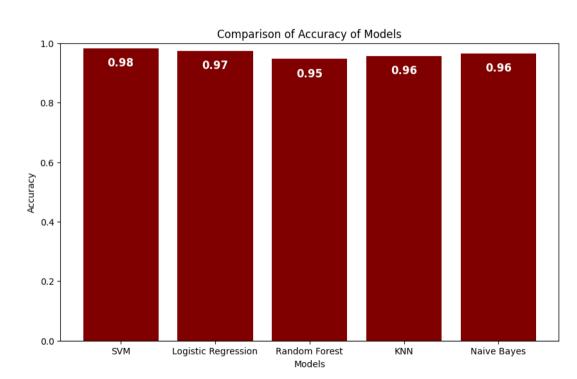
Classification	n Report of	SVM:		
	precision	recall.	f1-score	support
0	1.000000	0.953488	0.976190	43.000000
1	0.972603	1.000000	0.986111	71.000000
accuracy	0.982456	0.982456	0.982456	0.982456
macro avg	0.986301	0.976744	0.981151	114.000000
weighted avg	0.982937	0.982456	0.982369	114.000000

Classification	Report of	Logistic	Regression	
	precision	recall	f1-score	support
0	0.976190	0.953488	0.964706	43.000000
1	0.972222	0.985915	0.979021	71.000000
accuracy	0.973684	0.973684	0.973684	0.973684
macro avg	0.974206	0.969702	0.971863	114.000000
weighted avg	0.973719	0.973684	0.973621	114.000000

Classification	Report of	Random Fo	rest;	
	precision	recall	f1-score	support
0	0.930233	0.930233	0.930233	43.000000
1	0.957746	0.957746	0.957746	71.000000
accuracy	0.947368	0.947368	0.947368	0.947368
macro avg	0.943990	0.943990	0.943990	114.000000
weighted avg	0.947368	0.947368	0.947368	114.000000

Classificatio	n Report of	KNN:		
	precision	recall	f1-score	support
0	0.931818	0.953488	0.942529	43.00000
- 1	0.971429	0.957746	0.964539	71.00000
accuracy	0.956140	0.956140	0.956140	0.95614
macro avg	0.951623	0.955617	0.953534	114.00000
weighted avg	0.956488	0.956140	0.956237	114.00000

Classificatio	n Report of	Naive Bay	yes:	
	precision	recall	f1-score	support
0	0.975610	0.930233	0.952381	43.000000
-1	0.958904	0.985915	0.972222	71.000000
accuracy	0.964912	0.964912	0.964912	0.964912
macro avg	0.967257	0.958074	0.962302	114.000000
weighted avg	0.985205	0.964912	0.964738	114.000000

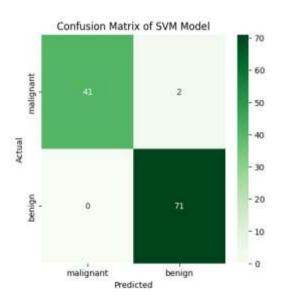


Based on the evaluation results of the five models, it was found that the Support Vector Machine has a higher accuracy compared to the other four models, namely Logistic Regression, Random Forest, KNN, and Naive Bayes. Therefore, the next step is to conduct a misclassification analysis and overfitting underfitting check.

BEST MODEL ANALYSIS

BEST MODEL ANALYSIS

Training data accuracy: 98.68% Testing data accuracy: 97.37%



The confusion matrix shows that the SVM model performs well in classifying breast cancer cases. It correctly identifies 41 malignant and 71 benign cases, with only 2 false negatives and no false positives. The high training (98.68%) and testing (97.37%) accuracy indicate that the model generalizes well without signs of overfitting or underfitting.

BY: NAFTALIA SOPHIANA PURBA

CONCLUSION

Breast cancer classification was performed using five machine learning models: SVM, Logistic Regression, Random Forest, KNN, and Naïve Bayes. SVM achieved the highest accuracy (98.25%), making it the best model.

Misclassification analysis showed minimal errors (2 false negatives, 0 false positives). The training (98.68%) and testing (97.37%) accuracies were close, indicating no overfitting or underfitting.

CONTACT

Email : naftaliasphn94@gmail.com

GitHub: https://github.com/naftaliasphna



BREAST CANCER CLASSIFICATION: MACHINE LEARNING APPROACH