Applications of Machine Learning in the analysis of breast cancer

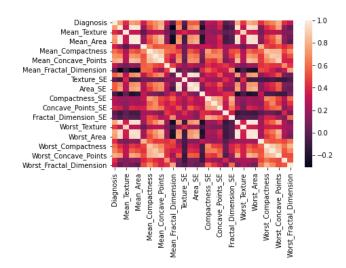
By Ultan Kearns & Liam Millar

Introduction

- When starting this project we investigated many datasets finally settling on https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/
- This dataset is from the University of Wisconsin a well-known institute of higher learning in the USA
- The main objective of our project was to use Machine Learning models to accurately predict the presence / absence of breast cancer based on features in the data
- We used a combination of both Supervised and Unsupervised learning in this project to train our models

Cleaning and analyzing the data

- First thing we did when starting this project
- We noticed diagnosis had M for malignant and B for benign we decided to replace these with 1 and 0 respectively
- Did this so that we could perform numerical operations easier
- We started by showing a heatmap of our correlations as you can see on the right hand side
- Did this to analyze relations in the data and to see which features were highly correlated with our diagnosis
- We also played around with the training / test set ratios and finally settled on a 70 / 30 split



	piagrocia e	NILNEL P	SPLTHERY MAN	Ordister	Post, 876	PARL SHIPTING	No. CHARLES	No. Desta	NAUTHOR MAN	man, sympto	MACHINE SHOULD	ratios, st	meters, N	NUTSHIELD, H	818,38	SACTOCK, III	CROKETSHILE.	overvitum.	DEPOSIT NAME	10,000	metal, rewester, re-y	MYS. MESS
Import	1,00000	Atletia	0.45915	874918	d Indee	*100744	0.594000	1 14070	0.77259	+3000	4 45630	F36143	941100	0.074326	10000	4.055741	1,9996	0,51608	0.00(07)	0.000007	0.00000	4,7980
Many Radice	atomi	1.00000	126707	sarw.	13040	Emerio.	93900	1 10070	0.01900	0.12800	43000	0.006761	0.80310	148311	12000	4.3(10)	4.37190	0.10000	0.54(%)	4007404	4.0000	6,870.00
Mari, Sottice	6.42210	436eE	100000	0.00025	1340	00404	100000	0,00166	63890	9 8000 W	4.00%	stron	0.MIM	325604	same:	4457639	6 HE2H	0.10394	410000	466781	14796	XIDE
ban, Personal	1.5600	110716	9.39215	1,00000	110'00	83664	1.0050	8.990.91	0,4607	0.0000	4,3954	C.92894	41970	3.753413	8 96519	4.161638	4,9429	0.00574	5.076040	4.503408	8.89681	6.815504
Merc Area	67169	110048	0 MON.	100754	1 310004	11740	0.500(9)	147004	C42(+9	0 15000	42300	6736450	400001	0727140	11901	4.6650	63360	0.179033	6.99716	410401	4 10000	contra
us, Secotions	630000	0.99790	4000	10551	11140	1,00000	12400	139411	05614	9 14000	13000	0.00000	000000	1297904	12199	1.502409	11000	0.214800	0.5900	6237578	1,794	121901
m,Comparison	£ 50400	Heck	0.000k	9.90786	11000	00004	1 (4000)	110w	0 0 0 0 0 0	9.9660	1 (1000)	COHM	007004	15000	1.000	0.133640	A.72960	110001	come	12044	1466	85400
man, Committee	C40710	446596	4.00 heb	0.7062+5	14750	0.59(74)	02010	1 30000	031340	0.40090	£30254	0.687364	0.104636	04F4200	0.00047	6 167600	667884	14/4/10	6.675167	4.301025	1.40030	1 1000
Linear June	4.7758W	14400	17000	110075	163400	1964	0.60400	63040	1 00000	0.4676	4.07989	ATENO	688301	9754624	ATHORY:	1.04000	14400	1 1000	0.00000	£192W	1,000ac	66090
han Symmetry	C3600	6 159CH	440004	110700	110006	19900	2.50000	6 46000	640%	1.00000	6,47960	KHIMN	0100214	110450	63940	£1000	0.0040%	8.361801	110/30	A.CWIN	11000	1 10340
Freist, Stemmer	A 69000	4 30 5000	A 1202%	.02016	427601	ammer.		63077	0.51768	0.07500	1	1100001	0.195427	30000	25700	1.3000	*8115	14430	137900	1.099	F#1000	ASMEN
Status, SE	6 603409	14600	470768	17084	8 75 MW	110000	2 (2000)	9900	67947	0 55866	1,000	1.000000	43944	9.966501	1345791	0.1477.0	11910	0.30380	110010	63044	8.700MW	43690
Index, St	serne.	AMORE	9.30206	20000	-01000	2,00011	1000	2 (160)	60600	2 1607	3 7947	6220018	1.000000	628810	E 00000	120100	43000	0311988	5,2694	1,000	1,1000	Asses
Samuel St.	0.01000	14800	62781K	1000	17716	828744	0.66600	+6100	0.73403	0.34000	1,000,00	1100001	4.00041	1-960000	ERRIT	3 (1)461	0.00110	9.348mc	61600	8.303604	# 350°000	8.730171
814,00	-	0.79mm	9,71971	0.7661)	ARREST	17040	9.8000	0.00000	071000	0.000	1070	4349791	91940	1908917	120000	100011	4,57999	0.01100	8.66747	0.00004	8.00070	A Person
Si, oresissee	ARREST	A SERVET	40000	211100	Alerta	STEELER .	9 (1241)	211191	0.00000	0.18030	1 1	GREEN	4.76160	611140	name to	1,00000	6.16316	630790	6.00000	6.69671	2.380ms	A30001
manines 70	12000	12796	0.98276	1000	6,000	130001	21286	19700	1 000	1,0000	1,000%	1279.00	120111	10070	surren.	134940	1,00000	1200	67070	14000	1,0000	12100
Consents, NE	\$ 2180R	0.798000	0.15500	20001	41797	525.000	2.00000	1000	0.19827	0.00000	0.00000	0.007991	0.077909	3.546544	1.01000	12220	A.76560	1.000000	677879	1 70004	8.76800	1 1000
nes, form, 30	AMERICA	0.24(%)	2 12800	827096	4.387%	1390	240000	0 artiror	0.70000	0.364.00	1,7300	CMIN	0.000064	A486754	0.460797	6.DITTR	6.75/910	171MB	1	8.09250	1.040	1.13000
Symmetry, SE	CONCT	40000	007752	-1000	0990	623979	120100	1293	0.1979	0.000	1.2900	622760	0.0000	22004	E2074	0.01027	14000	0.99900	6.9675	120009	1,000	11000
M. (NewFest), M.	14000	+++++	1000	0.00068	4 50000	107944	9,465400	14000	0,0004	0.0000	1,07070	1.15/WH	2771444	0,07349	0.19675	2,0004	6.866(1)	a Year	5.00429	1.000%	1,0000	45000
Minist Parlies	8.7780W	11010	17046	0.07534	127912	5272429	11400	2 40000	11010	0.10300	4,74621	5747367	C 200410	0798171	6.76003	a.lmmr	0.7 mires	0.00000	9.09000	4.534	-0.00079	1 stems
Marrel, Tanton	6 ellied	9.109HC	10000	9:000K	121100	10000	12000	6,00000	6289	+000	4 1000	0.109403	EH140	3196035	#200%	4.0004	0.0000	10000	1.15004	490006	4.03407	13000
ent Permits	17000	1 90002	1.0000	10/240	1200	8,20087	13650	9.704076	0.000	0.2*****	4.95594	6747966	00000	2722740	6.796723	4.9790	12910	111100	15729	4 mms1	1000	19072
NOSE, Ann	£ 79000	11400	1004	nowite	6 delines	100mps	45000	1450	33003	0.0000	4,040	0.75040	-0.000014	9792569	emmi.	4.94(0)	17130	0.10004	11006	×100410	48000	AMERICA
og, Steamforese	6.4000	810700	\$ perils.	0.00724	0.13047	+10000	15000	the section of	0.4000	0.6060	14000	61479	-0.710061	9.126726	E 146125	12990	1.19600	111100	T 100011	6.00407	8 100168	6.0000
a Congestions	ASCION	+40mm	939WT	SHOW	9.40000	0.45566	13504	1748	0.0000	949000	1000	E114534	-0.000719	309404	£3384	414110	16600	SARTH	0.0540	6479629	0.00mm	6.40000
tore, Concessor	covert	6195×	139HZ	1:59474	15000	14450	130400	1,0000	47900	0.40000	6 20000	640738	-0.04(4)5	5.840000	14000	414204	2.666701	0.4558W	450m	6.65588	8.96579	157095
Circum Paris	4 terms	# TSET#	5.565ua	a trainers	1700	0.48803	0.00000	4 MARCH	danie	0.000	0.00%	-	-6100166	3-556.504	1070	41000	6 amount	0.46780	1094	10700	4,000-0	A.76000
test Semestry	44000	4.00w/K	a degree	0.00044	1849	11000	0.50600	1 miles	0.000	1400	17000	E147676	-0.0000	3100040	1000	4 H000W	8.27 6627	41000	1.500	4.478625	8.11700	675006

Models Used

Linear Regression - to predict and analyze the correlation of 2 features in our dataset

K Means - This is an unsupervised learning technique creating clusters of data based on a centroid

KNN - This is a clustering technique which analyzes data nearest to other data to predict a diagnosis

Naïve Bayes - This is a technique used to predict cancer by taking features of our dataset and using the same weight for each -> Assumes data has same effect on output hence Naïve

Models Used Continued

Decision Trees - Which were used to predict diagnosis by making a tree of features in our dataset which will either lead to positive or negative diagnosis based on their values

Random Forest - Creates multiple trees then merges the best models to predict our diagnois

PCA (Principle Component Analysis) unsupervised dimensionality-reduction which reduced features in our dataset by removing similar features which were highly correlated

Analysis of our models

Linear Regression was used to analyze the strength of the relationship between certain features and our Diagnosis - Also used Cramérs V

Had good results we could see which features were positively correlated with a diagnosis value and the strength of this correlation

K Means - Performed fairly well with 86% accuracy

KNN - Performed better than KNN with 91% accuracy

Gaussian Naïve Bayes - without smoothing or scaling had 90% accuracy on testset

Analysis of our models

Continued

Gaussian Naïve Bayes with scaling had around a 91% accuracy rate - not much improvement

Decision Tree model had an accuracy of 93%

Random forest had around the same degree of accuracy

Entropy Forest had 90% accuracy

PCA Gini Tree had an accuracy of 0.94%

PCA Random forest had the best accuracy of all our models at 97%

Final Results Table

- Here we can see our final results table
- Notice which models performed correctly and which didn't
- Trial and error process it took time finding the right ratio of the training / test sets
- Also it took time to analyze the models and determine how we could get the best performance from them

A. Accuracy of Models

Method Description	Accuracy
PCA Gini Random Forest:	0.98%
cross validation gini random forest:	0.96%
Standard trained gini random forest:	0.95%
cross validation entropy random forest:	0.95%
PCA Gini Tree:	0.94%
Standard trained entropy random forest:	0.94%
Standard trained gini decision tree:	0.94%
cross validation entropy decision tree:	0.92%
Navie Bayes - with scaling:	0.91%
Standard trained entropy decision tree:	0.91%
cross validation gini decision tree:	0.91%
K Neighbour:	0.91%
K Means:	0.86%

Conclusion

From our study we we determined PCA Random Forest had the best accuracy when predicting the diagnosis

We also learned the limitations of machine learning in healthcare - should be used as an assistant not an expert as even the best trained models can yield false predictions

We learned the importance of data cleaning and analysis

We learned which models worked on our dataset and which didn't - it was a fairly small set of only 570 values!

Questions STIONS