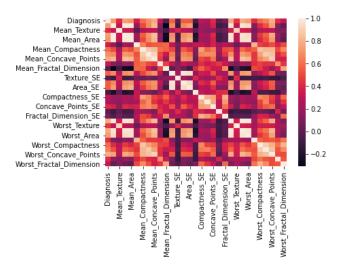
## Applications of Machine Learning in the analysis of breast cancer

By Ultan Kearns & Liam Millar

### Introduction

- When starting this project we investigated many datasets finally settling on <a href="https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/">https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/</a>
- 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 was our first task
- 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



		man, pecial m	MILTERY MIRTH	rische	Mary Area	Aut. Soctions	PER EMPETERS	Man, circle (5)	NW.CHORLEGES		met, rocket, merelier		10000,00	NUMBER 16	ANN, TE	SENSENHAL DE	OWNERSON, N		36 CROHLMISS,30	Specify St. 1	RAISHNINE STATE	
Dispose	1,00000	0.734900	6.49000 0	14979	0.714040	0.329701	0.00409	64070	8.775596	1,9600	-0.99500	8.900×05	6.01705	1.0100	0.50400	4.50794	3,0404	1 129	OR. CHAPTE	940341	0.00000	1,75000
Mon, Salice	0.737921	1,000000	0.366901 1	MALE	1760401	4 Names	61903	1 84576	1.0000	11000	-0.00360	1 607	4 96000	13000	274900	4.30(80)	0,0194	0.100	(34(f)) (m)	000104	4.0000	137793
Most, Sortice	6.00(0)	0.200407	100000 (	Hes	0.54(10)	444901	622994	6,307686	13040	143516	-912554	10076	0.00059	6290x	425979	410'00	1965	0.40	GA: 0.15000I	00190	68296	17092
Man / Verbooks	1.5650	2:92150	0.99ES	. 000000	1800	£20000	630755	0.700219	11075	1 107/00	0.25090	177000	4.99092	17045	it factors	4 10109	1,0425	6.99	P1 #31000F	40040	1,0000	8.97554
Mars, Area	£7940m	(1004))	0.040100 0	MONTH.	1 (0000)	0.07400	- comm	60((4)	130000	1 10000	425841	1790	4.9000	672746	HHDS	419010	1260	8490	PE CMIN	406432	4 0000	1000
New, Incohes	1.100%	0.105741	444011 1	264's	0.174003	1.00000	0.640624	0.004716	6.54De	11000	0.001507	0.00000	13969	4,36741	9.25+mm	11000	.0 0000	6296	00000	9.221679	1,27mm	62mox
den Conseitem	119401	HHOI-	6399H (	HOME	+5900	0.040001	1 00000	68744	1,0000	15000	#3660s	1500	60000w	6 (6400)	94009	+120H	47790	1 199	* 10000	03(144)	14040	£3400
Man, Committe	66070	6.665%	£38566 (	1606	0479346	8,000710	65744	4 3400000	17040	6.4055	936573	104704	1.0408	1500	1400047	4 10000	14709	1 100	661540	0.00+125	14000	E-6000/K
ner, forener, Perm	£.770000	144600	47mm (	MARCH	15000	2 04596	0.0000	610 MI	10000	14050	0.177900	870997	18005	8.79904	a retain	114598	0.45400		1010H	019790	1,7950	9.000000
Mon, Synmetry	£36000	0.101001	04001111	AFTER	0.150090	11000	£ 500 mm	6.40333	1.40%	-	0.075642	830060	11004	13454	+284CE	0 xx0xxx	0.0047	0.000	61 (1967)	440600	(100	0.00000
or Presid Discourse	41900	43000	0000	2900	.42989	1000	0.00004	4.8671	6.07000	1000	1,00000	0.00000	+ 3647	88802	Actions	4 34600	9 6627	140	EE 237000	427000	1000	43400
Subsa, M	FREIGH	64MCET	6271783 0	50,000	0.156430	6 909403	4.55will	6467394	9 700NT	a triberal	130000	1390000	6.2596	1900	934070	114700	42300	1 157	ET 271114	1223(4)	670906	£1690
Stellare, SE	0.613000	4 66500	0.00k2m	HERE	41001	FOREST	0.078834	0.09400	1000	1 1000	1000	1.7384A	1 00000	1,7000	9-120-020	0.79000	1,7611	400	0.000	Accept	10000	A STRUCK
Person, N	0.514130	1 (44)	039884 I	NOVE	0.027146	438790	6 66650	8,61400	\$79806	13800	9.09852	a battan	6.788k)	1 20000	NUMBER	413969	3 4613	1 1385	6A E100731	0.00004	630508	6750071
Ave. St.	13603	0.7880E	0219011 0	1900	000000	12290	0.0000	200047	8.74DK	12000	-0.079199	194720	0.0000	16397	1 000000	1.00000	9,200	130	0.665717	0.000339	8 1997	17090
Senten, S.	4.0000	4.30 (907	445500 1	BICK	0.0000	0.331am	0.02819	2 9794	0.00000	1 3000	13860	816729	1200	4 (0.00)	1000	1.00000	5 5000	627	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	0.010275	13000	4 2000
timestres, M	1,79490	92196	9.9609	PACEN	0.236477	9.80901	0.75898	5230	1000	4,0000	10101	1000	6,04000	14079	9.50100	0.0000	1,0000	170	H 179911	140011	6 89477	17000
Comarky, NI	621808	0.70000	0.00000	980	0.016073	921460	5.53816	20596	13607	1,0000	0.98800	110091	12/760	8.56mg	9300M	12000	37516	1.000	4.7716.01	0.30962	1760	6.760 pt
Court, Posts, 26	8.89875	13401	0.12800 1	1798	0.398799	0.0480	04000	2644	0.550000	120/06	0.07900	200M	1.23034	Keens	-	12000	97090	100	10000	0.046702	1643	13800
Spready, 55	1.0307	0.027600	441700 -	10000	0.004400	12397	0,300438	1.99(35)	A cores	1000	17760	1,01145	t-come	12000	4 20070	3.45025	2 6000	0.000	6346773	1.000001	1 1007	4.1004
alst (Newwood, M.	1.00007	44000	6427655 (	14490	-0-000000	12700	0.0040	0.40030	\$.340mm	1,7900	0.000(1)	KIDSSÓ	12044	1,30'00	9.19070	43604	0.00077	0.766	00 110400	COMM	1	-448079
MINISTRAL	6.77900	nation	0.1040	of those	0.079752	921069	Sizerer.	2 (000)	1.0000	1,1000	42601	1000	6.88400	a tarre	2.70mm	1200	927869	1781	11 110000	(0.10384)	4 6807	1.646000
Marri, Society	0.6560	11010	C01004 I	12980	0.290000	1000	6,00047	0.30000	1290W	1,6546	4.55000	1:10055	14549	3.9502	12059	41209	2770	+90	100000	4386N	42047	4.30mm
Brint Semester	a rices	1,94400	1969	19500	11206434	0.33580	0.597687	0.701	8.896/90	1,77504	0.555150	107706	6.F30H	1500	a mainte	4.5000	4,050	1 100	68 0.071790	-0.009461	1000	8.96579
Street, Ann	4.736600	6.642500	0.335546 1	1644 (DI	OMINI	0.30464	45000	14710	6.61004	3.0500	July July 1	3275045	4.8800	476265	0.00360	4 Hell	10409	1 110	64 635686	-0.00044)	41990	£36904
Maral, Streetholese	6.4000	11010	0-0418E - 1	MOS	0.136477	HARM	6.00401	1 4001	1.000	1.6947	0.00000	11499	6 9300	1905	0.94000	4,7976	11992	1 100	00 0.10001	0.0004451	6 125.00	12300
not, Corporate	£1000001	14000	4294477 (	4 ties	0.813300	4.400E	64004	674800	1,6957	1.0539	11000	130004	43875	63901	9.50mm	40000	1000	140	75. 0429401	HITMIN	627040	14000
West, Concepts	0.044007	640004	938HET 1	- MARCIN	0.543580	8.44525	coverty	0.00000	67900	1.0004	0.00000	140770	43655	1 40000	147590	41405#	cours	0.000	00 0143000	046500	1997	1576765
est, Compan, Forms	43000	A THAT'S	630040 1	. Nacri	0.793066	- AMACIE	1000	6 815 125	8 800000	6,4980	1,000	2100mi	4 0000	1.10054	H1768	41000	0.40006	0.600	10 001004	0.007589	6.00mm	6.76006
Mark Scientis	1000	E Mines	CMINI I	- Chicago	1 the ext	A ACRES	0.00001	1 44711	Lenes	10000	3.50000	114000	4.8876	A MOVE	à mene	A SHARE	4.7766	110	W 0.000	0.45000	6.45540	17900

### 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

**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** - Extended the decision trees to help remove the over fitting of the individual decision trees by taking the tallied vote of different data/decision tree configurations

**PCA (Principle Component Analysis)** - unsupervised dimensionality-reduction which reduced features in our dataset by removing similar features which were highly correlated, while still retaining as much pertinent information as possible.

# 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 test set

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

Most of our **Decision Tree** and **Random Forest** models had a had high degree of accuracy. However with our smaller dataset we must consider the possibility of **over fitting**.

**Entropy Forest had 90% accuracy** 

PCA Gini Tree had an accuracy of 94%

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

### Final Results Table And Conclusion

- 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
- From our study 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**!

### A. Accuracy of Models

Method Description	Accuracy
PCA Gini Random Forest:	0.98%
cross validation gini rdm forest:	0.96%
Standard trained gini rdm forest:	0.95%
cross validation ent random forest:	0.95%
PCA Gini Tree:	0.94%
Standard trained ent random forest:	0.94%
Standard trained gini decision tree:	0.94%
cross validation ent decision tree:	0.92%
Gaussian Naive Bayes-with scaling:	0.93%
Gaussian Naive Bayes-without scaling:	0.92%
Standard trained ent decision tree:	0.91%
cross validation gini decision tree:	0.91%
K Neighbour:	0.90%
K Means:	0.86%
Gaussian Naive Bayes-with smoothing:	0.79%