BREAST CANCER L AND R CLASSIFICATION AND ANALYSIS USING MACHINE LEARNING TECHNIQUES

M.S APPLIED INFORMATICS – 2ND STUDY CYCLE
VYTAUTAS MAGNUS UNIVERSITY
KAUNAS, LITHUANIA

Research Work Done By Ravinthiran Partheepan

AGENDA

1. Machine Learning – Data Mining – Big Data Analytics – Data Scientist
2. Breast Cancer Prediction and Prognosis
3. Machine Learning Methods
4. Comparison of Machine Learning methods
5. Summary and Future Research

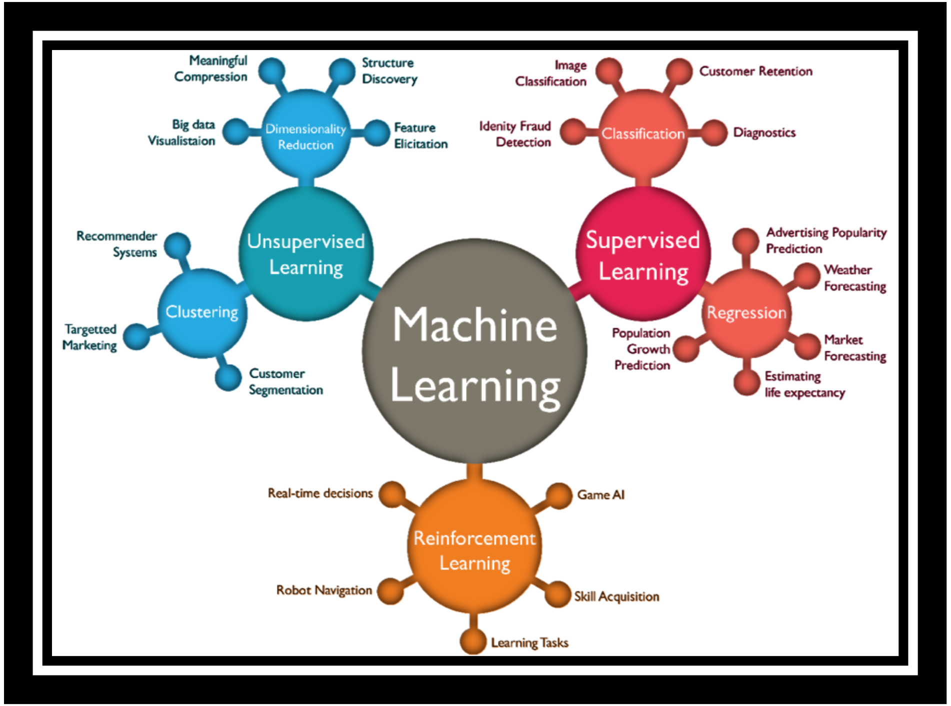
An overview about Breast Cancer

Breast cancer is the second cause of death among women. Early prediction of breast cancer will help with the survival of breast cancer patients. Data mining and machine learning have been widely used in the diagnosis of breast cancer and on the early detection of breast cancer. The aim of this research is to review the role of machine learning and data mining techniques in breast cancer detection and diagnosis. Most of these studies concentrated on diagnoses on breast cancer using WEKA tool.

*Keywords:- **Breast Cancer, Random Forest, Neural Network, Logistic Regression, Decision Tree, Machine Learning Approaches.***

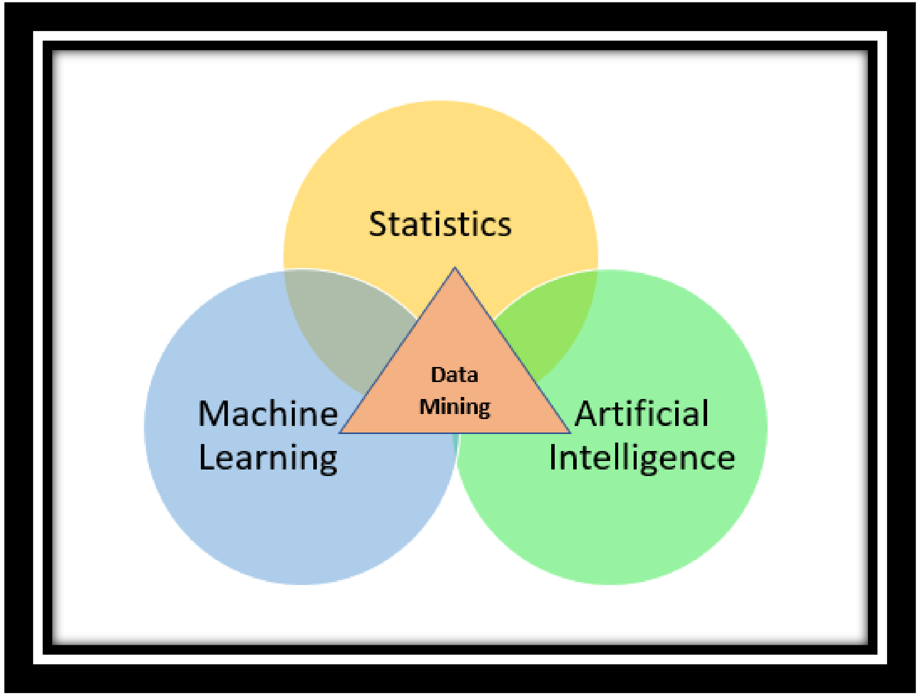
An overview about Machine Learning

1. Machine learning is
2. A branch of artificial intelligence
3. Employs a variety of statistical, probabilistic and optimization techniques
4. Allows computers to “learn” from past examples
5. Detect hard-to-discern pattern from large, noisy or complex data sets.”

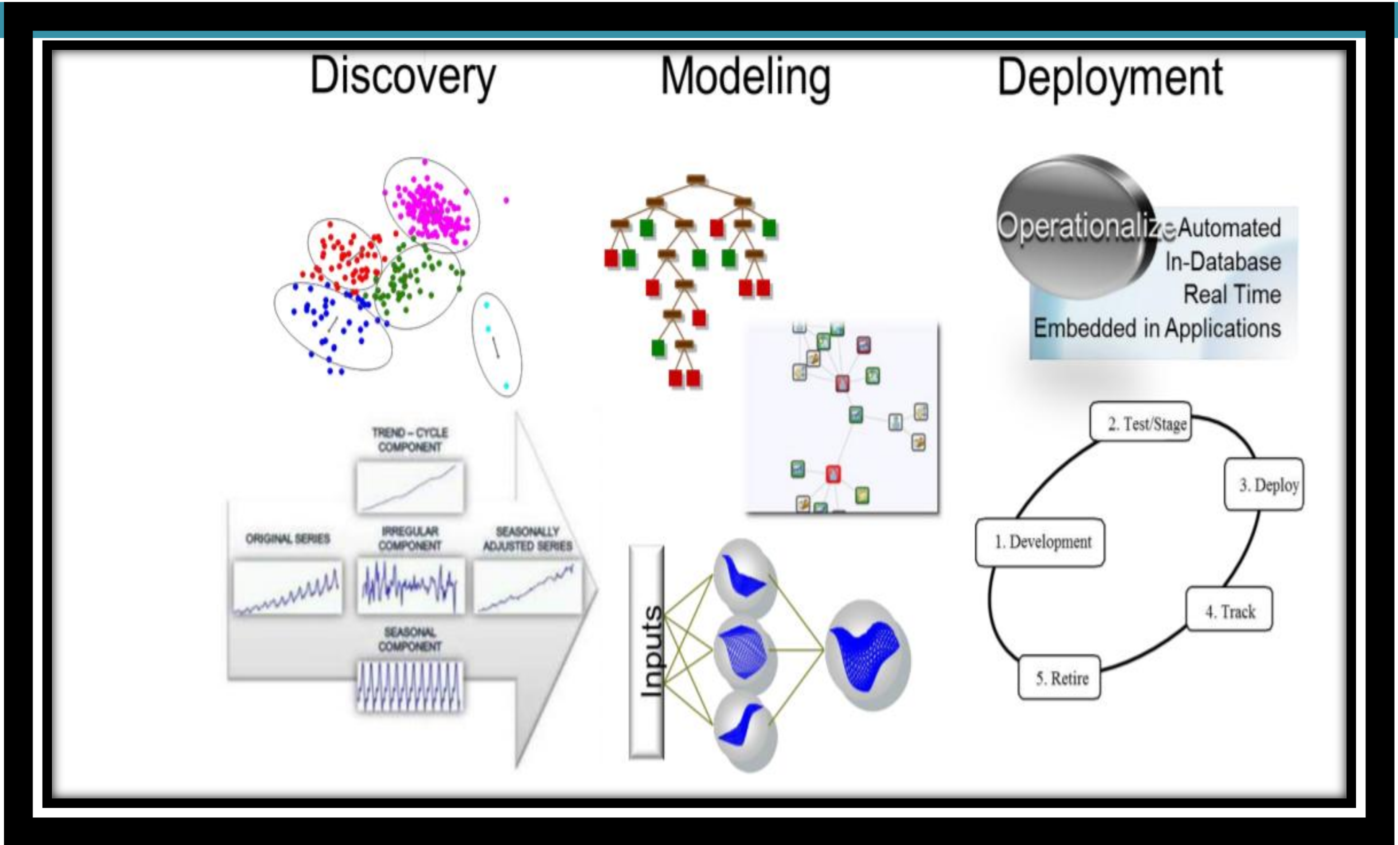


Data Mining VS Machine Learning

1. Machine Learning - Machine acquires knowledge from data
2. Data Mining – both Human & Machine together acquire Knowledge from data
3. Note that Data Mining and Machine Learning have been interchangeably used and appear to be overlapped in many ways.



Machine Learning Process



PREDICTION (DIAGNOSIS & PROGNOSIS)

Three cores of Breast Cancer Prediction and Prognosis:

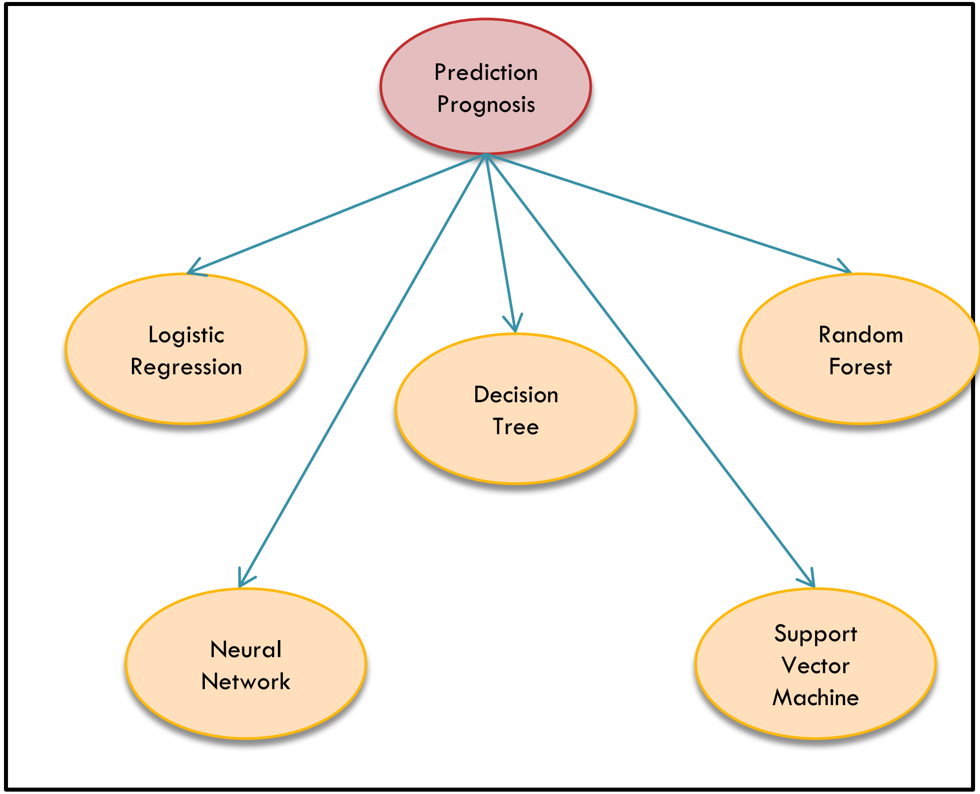
1. The prediction of breast cancer susceptibility – risk assessment prior to occurrence. (Diagnosis)
2. The prediction of breast cancer recurrence – likelihood of redeveloping (Prognosis)
3. The prediction of breast cancer survivability – life expectancy, survival, progression, tumor-drug sensitivity (Prognosis) The success of Prognosis prediction is dependent on the quality of the Diagnosis.

* **Note:-**

* **Prognosis** is the predicted outcome of a disease and the chances of recovery.

PREDICTION PROGNOSIS

1. Logistic Regression – predict the probability of the target event
2. Decision Tree – a segmentation of the data that is created by applying a series of simple rules.
3. Random Forest – multiple Decision Trees with random samples and random attributes. (ensemble method, hard to interpret)
4. Neural Networks – detecting complex nonlinear relationships in data
5. Support Vector Machines – construct a set of hyperplanes that maximize the margin between two classes for classification.



9 CATEGORICAL VARIABLES+ 5 CONTINUOUS VARIABLES

Categorical Variables:

1. Primary site code - SITE02V
2. Histology - HISTO2V
3. Behavior - BEHO2V
4. Grade – GRADE
5. Extension of disease – EOD10_EX
6. Lymph node involvement – EOD10_ND
7. Radiation – RADIATN
8. Stage of Cancer – D_AJCC_M 9. Site specific surgery code – SS_SURG

Continuous Variables:

1. Age at diagnosis – AGE_DX
2. Tumor size – EOD10_SZ
3. Number of positive nodes – EOD10_PN
4. Number of nodes examined – EOD10_NE
5. Number of primaries - NUMPRIMS

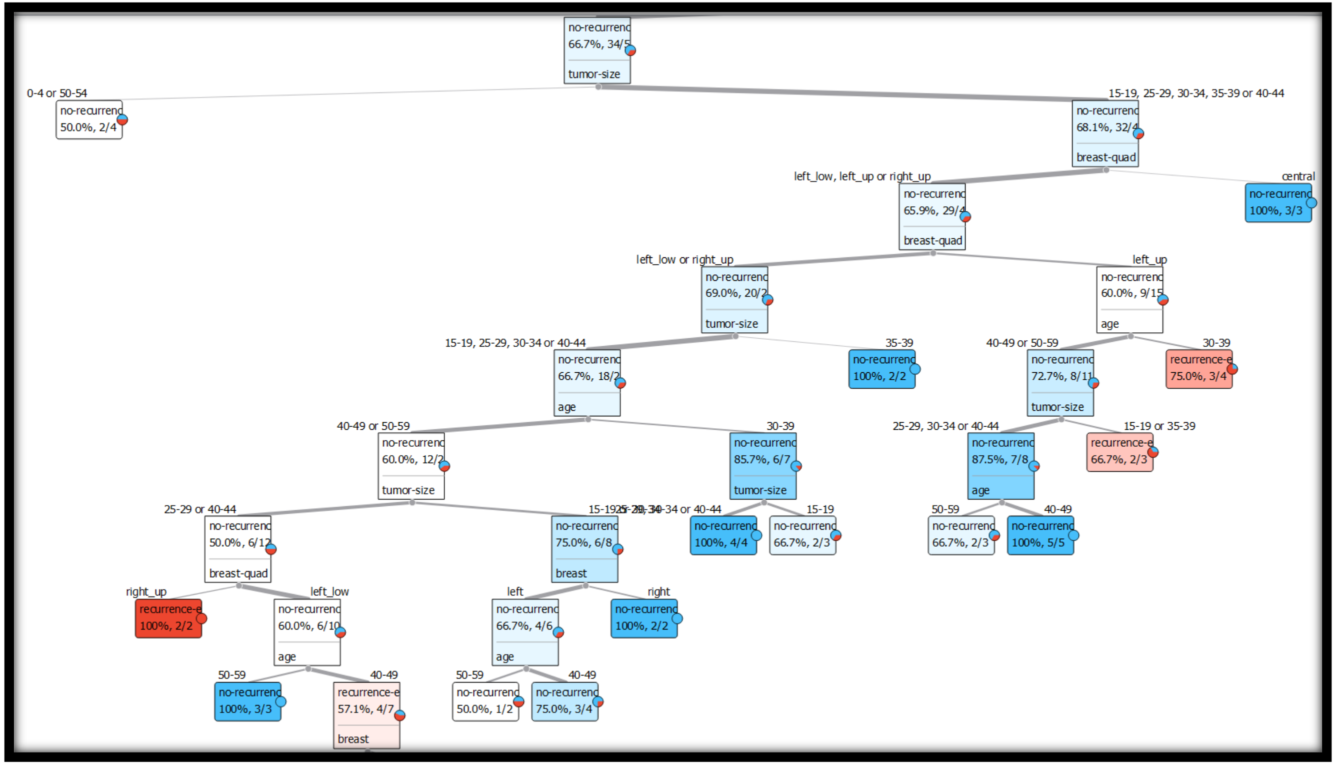
Random Forest

Loss Reduction Variable Importance

Variable	Number of Rules	Gini	OOB Gini	Valid Gini	Margin	OOB Margin	Valid Margin
EOD10_ND	293	0.032699	0.021873	0.025141	0.065399	0.043702	0.048640
EOD10_EX	348	0.030872	0.020317	0.021649	0.061745	0.040937	0.043966
AGE_DX	409	0.018423	0.011673	0.012378	0.036847	0.023983	0.025563
EOD10_PN	504	0.014227	0.009277	0.009330	0.028453	0.018727	0.019448
SS_SURG	510	0.014449	0.009249	0.010031	0.028897	0.018888	0.020369
BEHO2V	70	0.008151	0.005509	0.006317	0.016303	0.010958	0.012449
NUMPRIMS	400	0.006712	0.004345	0.004456	0.013423	0.008774	0.008990
RADIATN	312	0.003290	0.002034	0.002041	0.006580	0.004268	0.004445
HISTO2V	226	0.003175	0.001731	0.001841	0.006350	0.003892	0.004198
GRADE	368	0.002280	0.001200	0.001520	0.004560	0.002711	0.003272
EOD10_NE	111	0.001030	0.000571	0.000539	0.002060	0.001246	0.001251
EOD10_SZ	241	0.001060	0.000533	0.000562	0.002119	0.001239	0.001319
SITEO2V	89	0.000285	0.000055	0.000047	0.000570	0.000254	0.000265

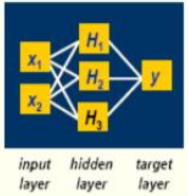
* Gini Ratio in **Random Forests** allow us to look at feature importances, which is the how much the **Gini Index** for a feature decreases at each split.

Decision Tree



Neural Network (Multi-Layer Perceptron)

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_{00} + \hat{w}_{01} \cdot H_1 + \hat{w}_{02} \cdot H_2 + \hat{w}_{03} \cdot H_3$$



$$H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11}\, x_1 + \hat{w}_{12}\, x_2)$$

$$H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21}\, x_1 + \hat{w}_{22}\, x_2)$$

$$H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31}\, x_1 + \hat{w}_{32}\, x_2)$$

N	Parameter	Estimate	Gradient Objective Function
1	AGE_DX_H11	0.220341	-0.001199
2	EOD10_NE_H11	0.056880	-0.000862
3	EOD10_PN_H11	-0.089935	0.006769
4	EOD10_SZ_H11	-0.059415	-0.003658
5	NUMPRIMS_H11	0.334654	-0.005225
6	AGE_DX_H12	-1.402353	0.001699
7	EOD10_NE_H12	0.182900	0.009414
8	EOD10_PN_H12	-0.367248	-0.014434
9	EOD10_SZ_H12	-0.047111	0.000252
10	NUMPRIMS_H12	-0.011302	0.007801
11	AGE_DX_H13	-0.359919	-0.000900
12	EOD10_NE_H13	0.053678	0.008494
13	EOD10_PN_H13	-0.392417	-0.009839
14	EOD10_SZ_H13	0.001905	-0.000032918
15	NUMPRIMS_H13	0.005708	0.008607
16	BEH02V2_H11	-0.665449	-0.004782
17	BEH02V2_H12	0.364894	-0.000932
18	BEH02V2_H13	-0.114147	-0.006538
19	EOD10_EX00_H11	-0.250320	-0.002365
20	EOD10_EX05_H11	-0.075988	-0.000211
21	EOD10_EX10_H11	-0.098221	0.002660
22	EOD10_EX20_H11	0.098889	-0.000299
23	EOD10_EX30_H11	-0.095367	-0.000405
24	EOD10_EX40_H11	0.034954	-0.000375
25	EOD10_EX50_H11	0.053517	-0.000465
26	EOD10_EX60_H11	0.003395	-0.000377
27	EOD10_EX70_H11	0.024721	-0.000577
28	EOD10_EX80_H11	-0.111060	-0.000376

1. A **multilayer perceptron** (MLP) **is a** deep, artificial neural network. … They are composed of an input **layer** to receive the signal, an output **layer** that makes a decision or prediction about the input, and in between those two, an arbitrary number of hidden **layers** that are the true computational engine of the MLP.
2. The **hidden layers'** job is to transform the inputs into something that the output **layer** can use. The output **layer** transforms the **hidden layer** activations into whatever scale you wanted your output to be on.
3. I have scaled hidden layer to 3.

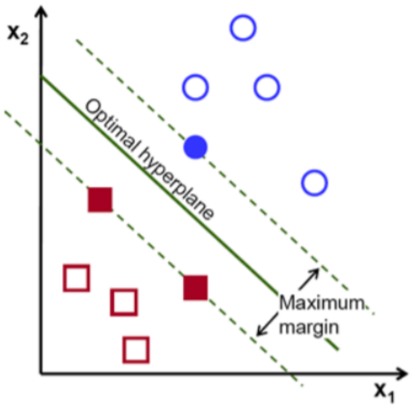
Logistic Regression Model

$$\log\left(\frac{p}{1-p}\right) = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	3.1809	0.8218	14.98	0.0001		24.069
AGE_DX		1	-0.0525	0.000665	6230.11	<.0001	-0.5576	0.949
EOD10_EX	00	1	2.1936	0.0742	874.76	<.0001		8.968
EOD10_EX	05	1	0.9504	0.3048	9.72	0.0018		2.587
EOD10_EX	10	1	0.9037	0.0673	180.56	<.0001		2.469
EOD10_EX	20	1	0.2767	0.0794	12.16	0.0005		1.319
EOD10_EX	30	1	0.0652	0.1062	0.38	0.5395		1.067
EOD10_EX	40	1	-0.0646	0.1508	0.18	0.6682		0.937
EOD10_EX	50	1	-0.2683	0.0992	7.32	0.0068		0.765
EOD10_EX	60	1	-0.8665	0.3668	5.58	0.0182		0.420
EOD10_EX	70	1	-0.3351	0.1231	7.41	0.0065		0.715
EOD10_EX	80	1	-1.1165	0.4970	5.05	0.0247		0.327
EOD10_EX	85	1	-2.1723	0.1034	441.32	<.0001		0.114
EOD10_ND	0	1	1.1205	0.0353	1010.35	<.0001		3.066
EOD10_ND	1	1	0.3740	0.0615	37.03	<.0001		1.453
EOD10_ND	2	1	0.1062	0.0531	4.00	0.0455		1.112
EOD10_ND	3	1	-0.1128	0.0708	2.53	0.1114		0.893
EOD10_ND	4	1	-0.3027	0.0572	28.00	<.0001		0.739
EOD10_ND	5	1	-0.4022	0.0596	45.61	<.0001		0.669
EOD10_ND	6	1	-0.0593	0.0370	2.58	0.1085		0.942
EOD10_ND	7	1	-0.1938	0.2161	0.80	0.3698		0.824
EOD10_ND	8	1	-1.2382	0.1677	54.50	<.0001		0.290
EOD10_NE		1	0.00141	0.000583	5.87	0.0154	0.0132	1.001
EOD10_PN		1	-0.00654	0.000395	274.06	<.0001	-0.1566	0.993
GRADE	1	1	0.4918	0.0262	351.36	<.0001		1.635
GRADE	2	1	0.1446	0.0176	67.62	<.0001		1.156
GRADE	3	1	-0.3504	0.0175	399.46	<.0001		0.704
GRADE	4	1	-0.2247	0.0379	35.20	<.0001		0.799
HISTO2V	8000	1	-0.5785	0.5741	1.02	0.3136		0.561
HISTO2V	8001	1	-0.4853	0.7946	0.37	0.5414		0.616
HISTO2V	8003	1	-3.8384	10.7535	0.13	0.7211		0.022
HISTO2V	8004	1	-0.3456	2.2863	0.02	0.8798		0.708
HISTO2V	8010	1	-0.9952	0.5622	3.13	0.0767		0.370
HISTO2V	8012	1	-0.1753	0.8585	0.04	0.8383		0.839

Logistic regression is a statistical **model** that in its basic form uses a **logistic** function to **model** a binary, dependent variable, although many more complex extensions exist. In **regression** analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a **logistic model** (a form of binary **regression**).

Support Vector Machine

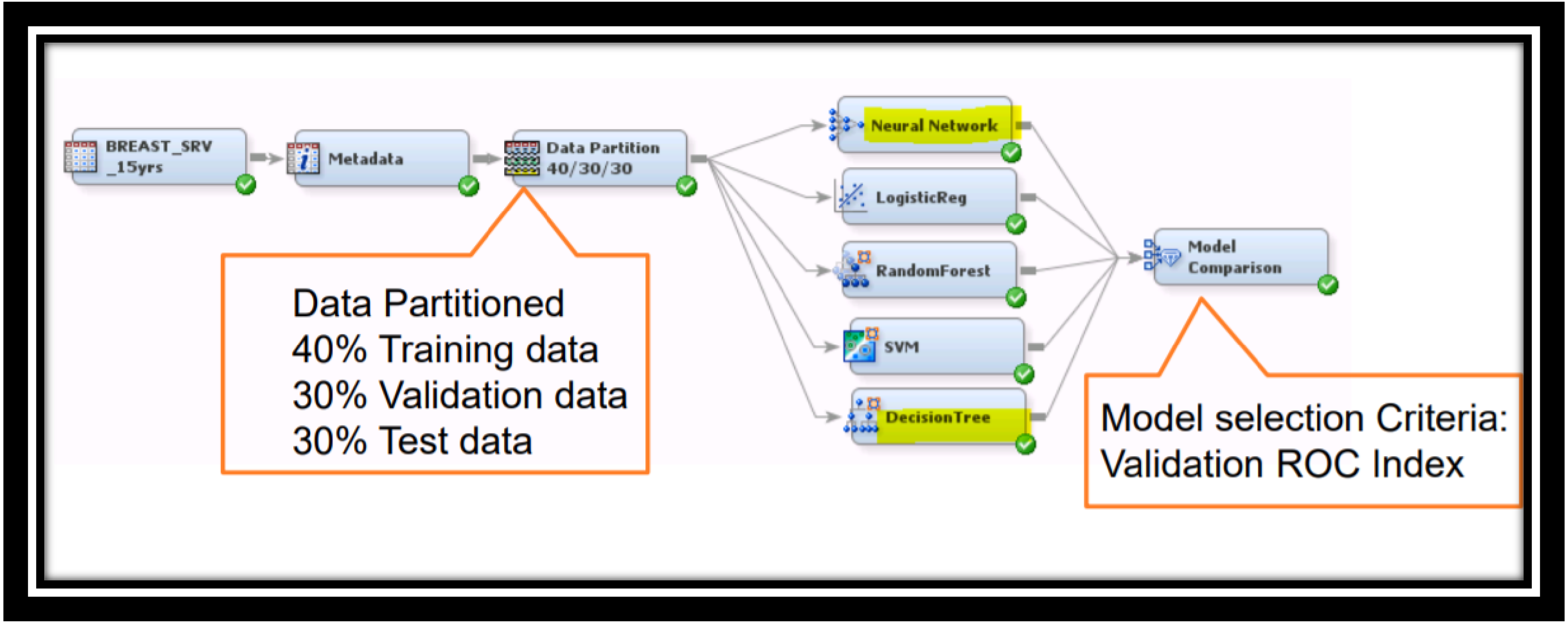


SVM Fit Statistics

Description	Train	Validation
Number of Observations Read ...	135486	.
Number of Observations Used ...	94839	40647
Number of Input Interval Variables ...	5	.
Number of Input Class Variables ...	8	.
Number of Input Class Variable Levels...	157	.
Number of Effects	13	.
Columns in Data Matrix	162	.
Inner Product of Weights	1365.071	.
Bias	-1.04983	.
Total Slack (Constraint Violations) ...	52030.57	.
Norm of Longest Vector	3.464102	.
Number of Support Vectors ...	94839	.
Number of Support Vectors on Margin ...	19	.
Maximum F	35.70521	.
Minimum F	-3.93071	.
Accuracy	0.764127	0.766477
Error	0.235873	0.233523
Sensitivity	0.892832	0.89564
Specificity	0.559527	0.561134

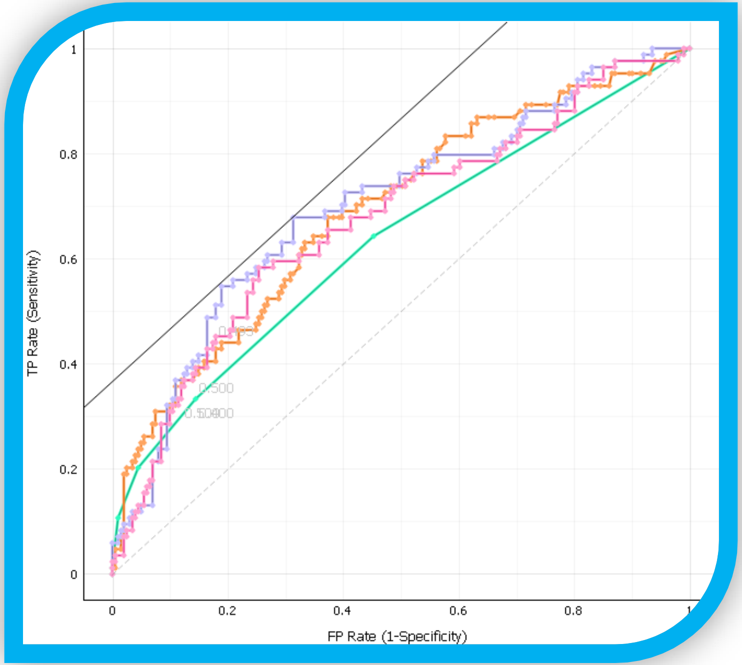
1. **SVM** supervised **machine** learning algorithm which can be **used for** classification or regression problems. It **uses** a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

Model Processing



Model Comaprison

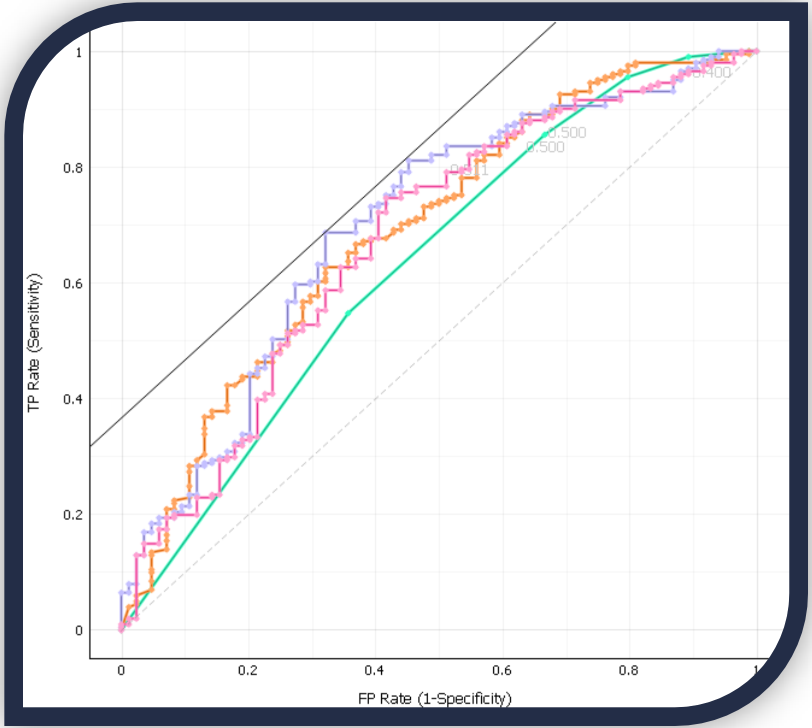
Method	AUC	CA	F1	Precision	Recall
Neural Network	0.645	0.734	0.682	0.719	0.734
Logistic Regression	0.679	0.724	0.701	0.700	0.724
Random Forest	0.694	0.717	0.701	0.696	0.717
Decision Tree	0.712	0.731	0.726	0.723	0.731



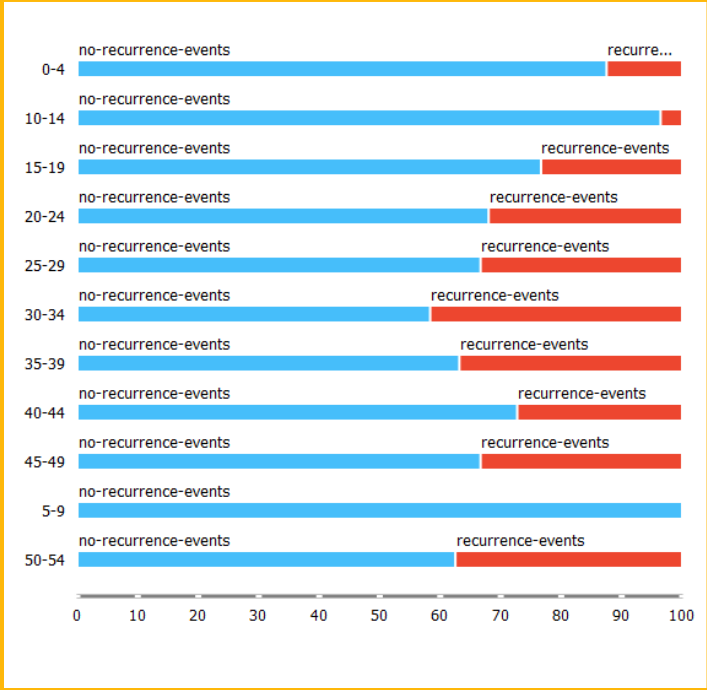
target	output	
1= survival	1=survival	true positive
0=die	1=survival	false postive
1=survival	0=die	false negative
0=die	0=die	true negative

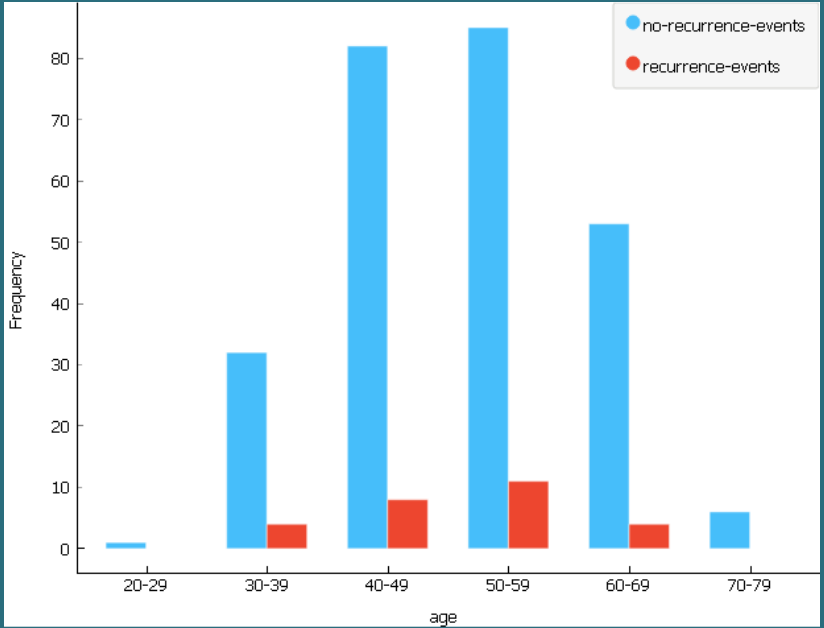
Classification Table

Event Classification Table						
Model Selection based on Validation data: ROC Index						
Model	Data Role	FALSE Negative	TRUE Negative	FALSE Positive	TRUE Positive	ROC Index
Neural Network	TRAIN	3955	13349	7578	29313	
Neural Network	VALIDATE	2980	9918	5777	21970	0.851
LogisticReg	TRAIN	4073	12384	8543	29195	
LogisticReg	VALIDATE	3066	9290	6405	21884	0.831
RandomForest	TRAIN	3664	12163	8764	29604	
RandomForest	VALIDATE	2783	9095	6600	22167	0.83
SVM	TRAIN	3523	11436	9491	29745	
SVM	VALIDATE	2640	8514	7181	22310	0.827
DecisionTree	TRAIN	6548	14344	6583	26720	
DecisionTree	VALIDATE	4899	10757	4938	20051	0.811



Statistical Report





Merits

1. Logistic Regression – Causal effect
2. Decision Tree – English rule, segmentation, variable selection, use both categorical and interval with missing values Random Forest – reduce overfitting
3. Neural Networks – nonlinear, local maximum
4. Support Vector Machine – nonlinear, global maximum

Conclusion

1. Machine Learning/Data mining is a key technique to automate Medical disease classification with much improved architecture.
2. Further tests and research are needed. Further specification: SVM (linear, polynomial, RBF, sigmoid kernel)
3. Methods: Clustering, segmentation, two stage modeling, cross validation Data: subsets (HER2+/-), different cancers, unstructured data
4. Architeture: HDFS Laser server, In-Memory statistics, Results Visualization

Thank you for listening

Do you have any questions?