6th Malaysia Statistics Conference
19 November 2018
Sasana Kijang, Bank Negara Malaysia

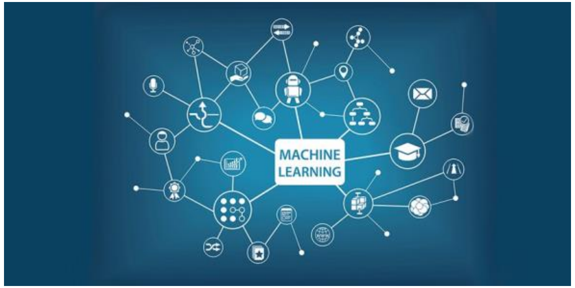
2018

Embracing Data Science and Analytics to Strengthen Evidence-Based Decision Making

Application of Big Data and Machine Learning in Bioinformatics

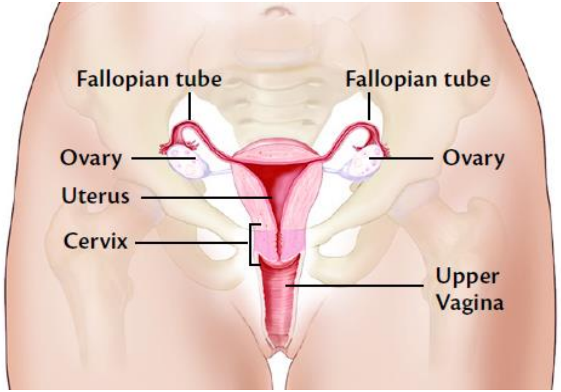
Azian Azamimi Abdullah
School of Mechatronic Engineering
Universiti Malaysia Perlis (UniMAP)

Introduction



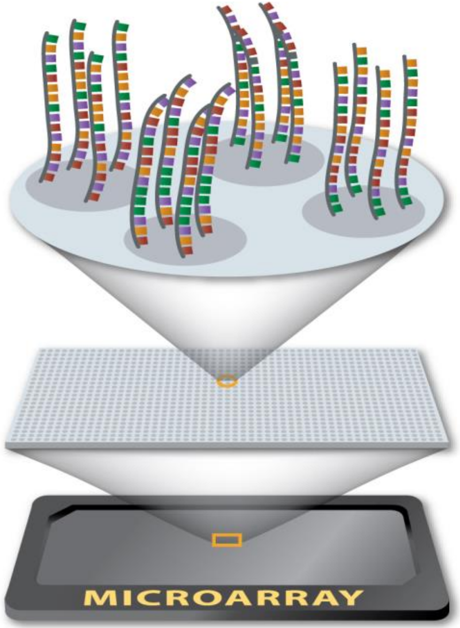
- Big data → data sets that are too large or complex for traditional data-processing or statistical application software
- Machine learning →Subfield of computer science, application of artificial intelligence (AI), provides systems the ability to automatically learn from experience
- Unsupervised and supervised learning
- Applications: Healthcare, **bioinformatics**, robotics, data security, financial market analysis, translation

Our Study



- Cervical Cancer
  - Second most common cancer
  - Caused by human papillomavirus (HPV)
  - Left untreated, cervical cancer developed
  - Lead to life-threatening but potentially curable

Microarray Gene Expression Profiling →To interpret and analyze the genes expression state in complementary DNA prepared from mRNA in which the hybridization is taking place on the array



Problem Statement

- Classification and detection of cervical cancer among large community has been challenging task
- Cervical cancer is detected by PAP Smear test which has limited sensitivity to detect the early development of cervical cancerous cell
- But by using gene expression profiling data, it has a better sensitivity to detect the early development of cervical cancer
- Hence, predictive model is important

Objectives

- To extract the important features from cervical cancer gene expression profiling data
- To classify groups or clusters of similar genes using unsupervised machine learning (ML)
- To develop predictive models for cervical cancer by using supervised machine learning (ML)

Scopes

- Dataset: Gynecologic Oncology Group Tissue Bank (PA, USA) & Kaggle database website
- Dataset containing gene expression profiling data in order to detect whether they are pre-cancerous or cancerous cervical cells
- Data is extracted to get the important features
- Unsupervised ML: hierarchical clustering & principal components analysis (PCA)
- Supervised ML: support vector machine (SVM) & Random Forest (RF)

Methodology

Dataset Description:

✓Gene expression profiling data

✓Tumor and matched normal samples

✓Raw read counts from the sequencing of microRNA

✓58 samples data with 714 features

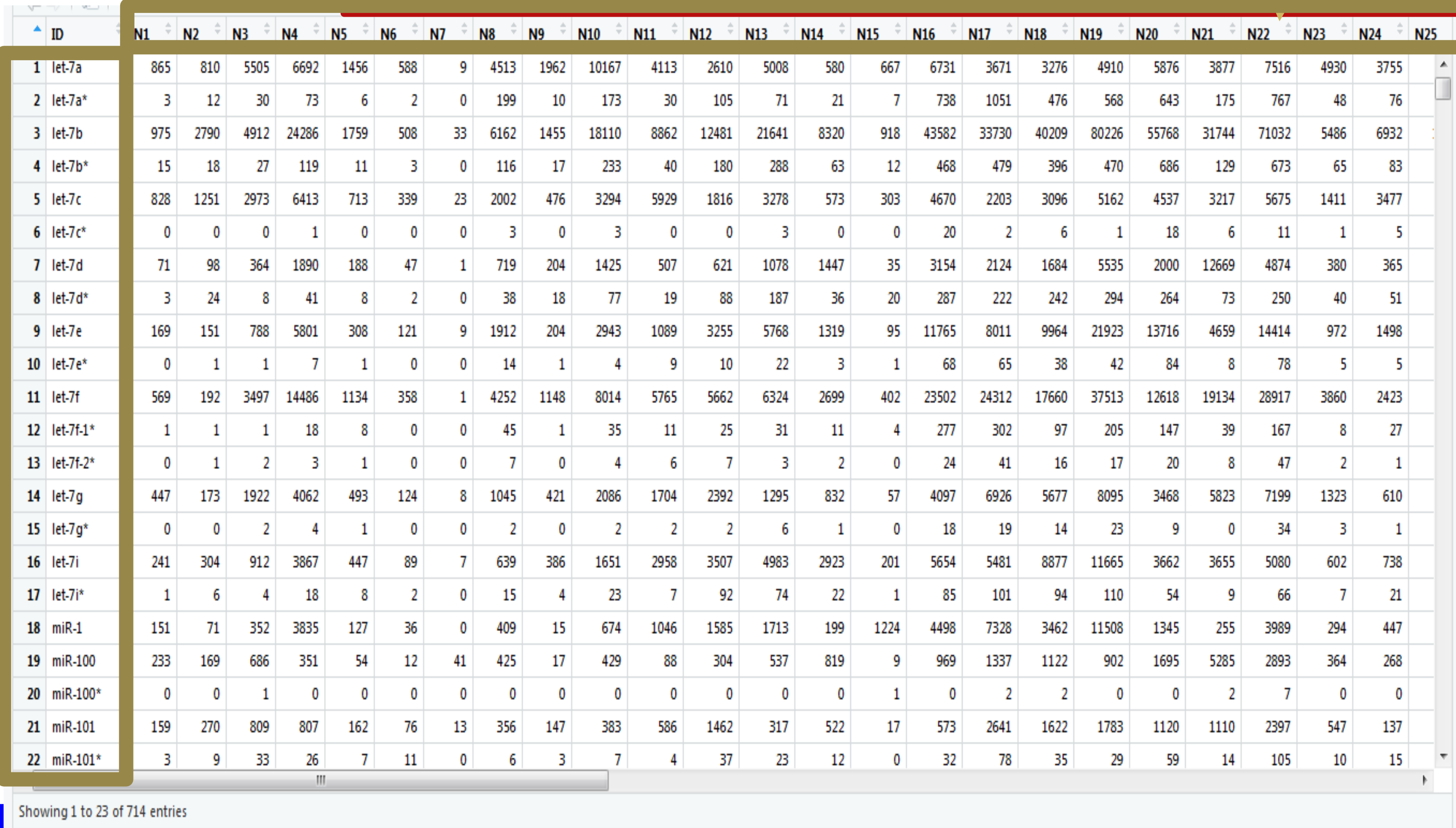
✓Row: microRNA features

✓Column: 29 Normal (N), 29 Tumor (T)

R statistical computing environment

Samples Data [Normal (N): 29, Tumor (T): 29]

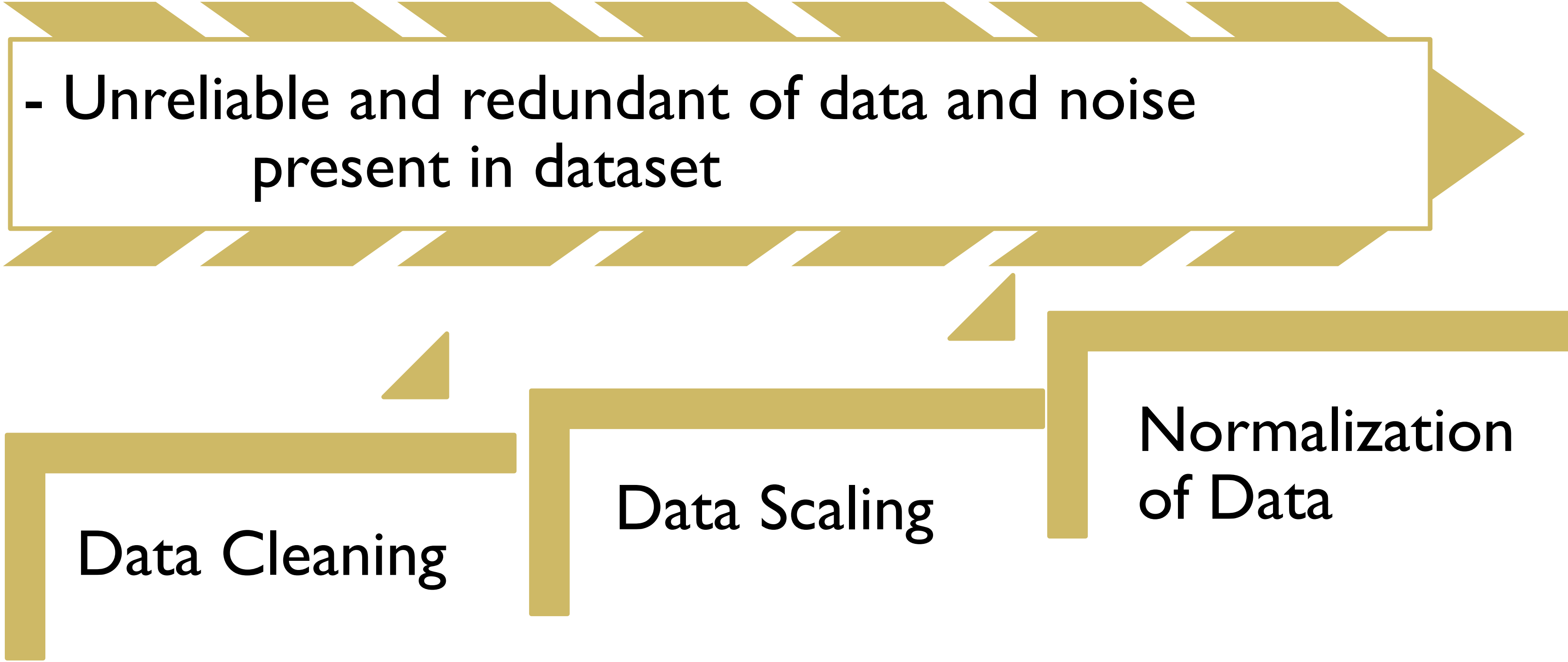
Features of Sequencing of microRNA



	ID	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	N16	N17	N18	N19	N20	N21	N22	N23	N24	N25
1	let-7a	865	810	5505	6692	1456	588	9	4513	1962	10167	4113	2610	5008	580	667	6731	3671	3276	4910	5876	3877	7516	4930	3755	
2	let-7a*	3	12	30	73	6	2	0	199	10	173	30	105	71	21	7	738	1051	476	568	643	175	767	48	76	
3	let-7b	975	2790	4912	24286	1759	508	33	6162	1455	18110	8862	12481	21641	8320	918	43582	33730	40209	80226	55768	31744	71032	5486	6932	
4	let-7b*	15	18	27	119	11	3	0	116	17	233	40	180	288	63	12	468	479	396	470	686	129	673	65	83	
5	let-7c	828	1251	2973	6413	713	339	23	2002	476	3294	5929	1816	3278	573	303	4670	2203	3096	5162	4537	3217	5675	1411	3477	
6	let-7c*	0	0	0	1	0	0	0	3	0	3	0	0	3	0	0	20	2	6	1	18	6	11	1	5	
7	let-7d	71	98	364	1890	188	47	1	719	204	1425	507	621	1078	1447	35	3154	2124	1684	5535	2000	12669	4874	380	365	
8	let-7d*	3	24	8	41	8	2	0	38	18	77	19	88	187	36	20	287	222	242	294	264	73	250	40	51	
9	let-7e	169	151	788	5801	308	121	9	1912	204	2943	1089	3255	5768	1319	95	11765	8011	9964	21923	13716	4659	14414	972	1498	
10	let-7e*	0	1	1	7	1	0	0	14	1	4	9	10	22	3	1	68	65	38	42	84	8	78	5	5	
11	let-7f	569	192	3497	14486	1134	358	1	4252	1148	8014	5765	5662	6324	2699	402	23502	24312	17660	37513	12618	19134	28917	3860	2423	
12	let-7f-1*	1	1	1	18	8	0	0	45	1	35	11	25	31	11	4	277	302	97	205	147	39	167	8	27	
13	let-7f-2*	0	1	2	3	1	0	0	7	0	4	6	7	3	2	0	24	41	16	17	20	8	47	2	1	
14	let-7g	447	173	1922	4062	493	124	8	1045	421	2086	1704	2392	1295	832	57	4097	6926	5677	8095	3468	5823	7199	1323	610	
15	let-7g*	0	0	2	4	1	0	0	2	0	2	2	2	6	1	0	18	19	14	23	9	0	34	3	1	
16	let-7i	241	304	912	3867	447	89	7	639	386	1651	2958	3507	4983	2923	201	5654	5481	8877	11665	3662	3655	5080	602	738	
17	let-7i*	1	6	4	18	8	2	0	15	4	23	7	92	74	22	1	85	101	94	110	54	9	66	7	21	
18	miR-1	151	71	352	3835	127	36	0	409	15	674	1046	1585	1713	199	1224	4498	7328	3462	11508	1345	255	3989	294	447	
19	miR-100	233	169	686	351	54	12	41	425	17	429	88	304	537	819	9	969	1337	1122	902	1695	5285	2893	364	268	
20	miR-100*	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	2	2	0	0	2	7	0	0	
21	miR-101	159	270	809	807	162	76	13	356	147	383	586	1462	317	522	17	573	2641	1622	1783	1120	1110	2397	547	137	
22	miR-101*	3	9	33	26	7	11	0	6	3	7	4	37	23	12	0	32	78	35	29	59	14	105	10	15	

Showing 1 to 23 of 714 entries

Data Pre-processing



- Unreliable and redundant of data and noise present in dataset

Data Cleaning

Data Scaling

Normalization of Data

Unsupervised Machine Learning

Hierarchical Clustering (Heatmap)

- Implemented to create the hierarchical
- By measuring the similarities between features of the gene expression profiles

Principle Component Analysis (PCA)

- Reduce the dimensionality
- Collapse the hundreds of features into a smaller set of principal components

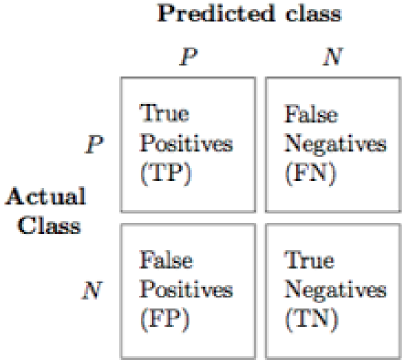
Supervised Machine Learning (SVM and RF)

- Support Vector Machine (SVM)
- Maximize the accuracy of the predictions while avoiding over-fit to the sample data
- Random Forest (RF)
- Ensemble learning method that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification)
- Training and testing the sample data (70% train, 30% test)

Evaluation of Model Performance

$$\% \, Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \; x \; 100$$

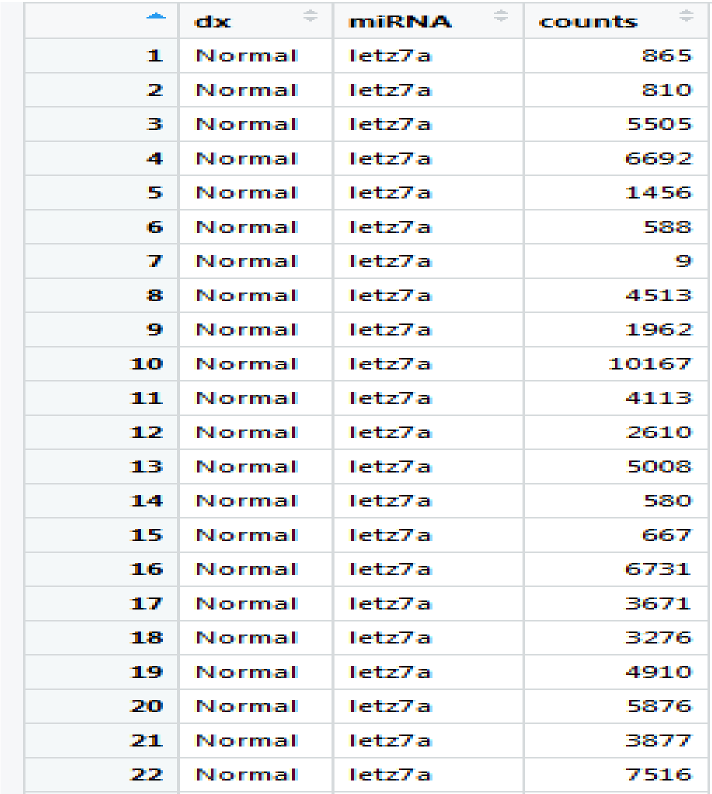
$$\% \, Sensitivity = \frac{TP}{TP + FN} \; x \; 100$$



		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

$$\% \, Specificity = \frac{TN}{TN+FP} \; x \; 100$$

Results and Discussion



	dx	miRNA	counts
1	Normal	letz7a	865
2	Normal	letz7a	810
3	Normal	letz7a	5505
4	Normal	letz7a	6692
5	Normal	letz7a	1456
6	Normal	letz7a	588
7	Normal	letz7a	9
8	Normal	letz7a	4513
9	Normal	letz7a	1962
10	Normal	letz7a	10167
11	Normal	letz7a	4113
12	Normal	letz7a	2610
13	Normal	letz7a	5008
14	Normal	letz7a	580
15	Normal	letz7a	667
16	Normal	letz7a	6731
17	Normal	letz7a	3671
18	Normal	letz7a	3276
19	Normal	letz7a	4910
20	Normal	letz7a	5876
21	Normal	letz7a	3877
22	Normal	letz7a	7516

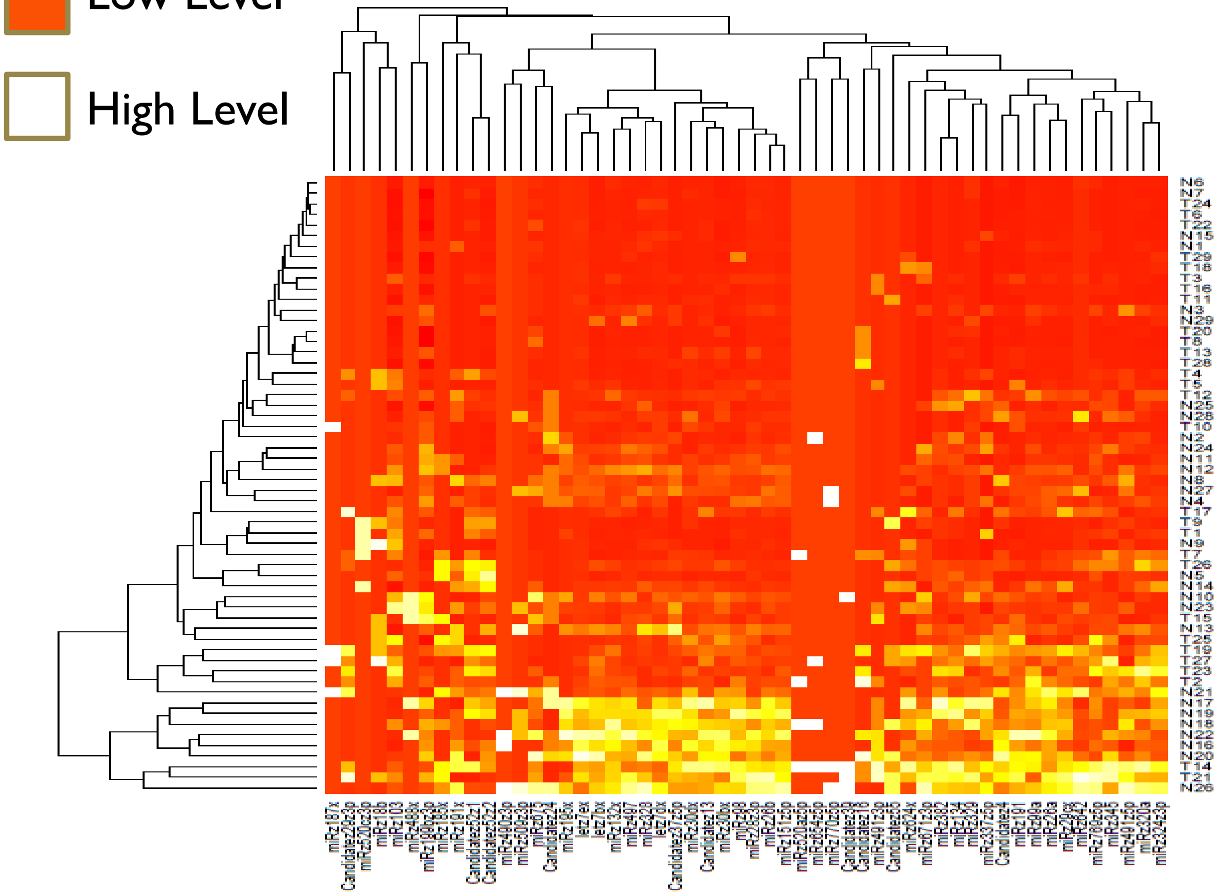
Data Pre-Processing:

- ✓ Unreliable data are removed
- ✓ Final dataset is used
- ✓ Also wide form of data change into long form
- ✓ Used in the next process

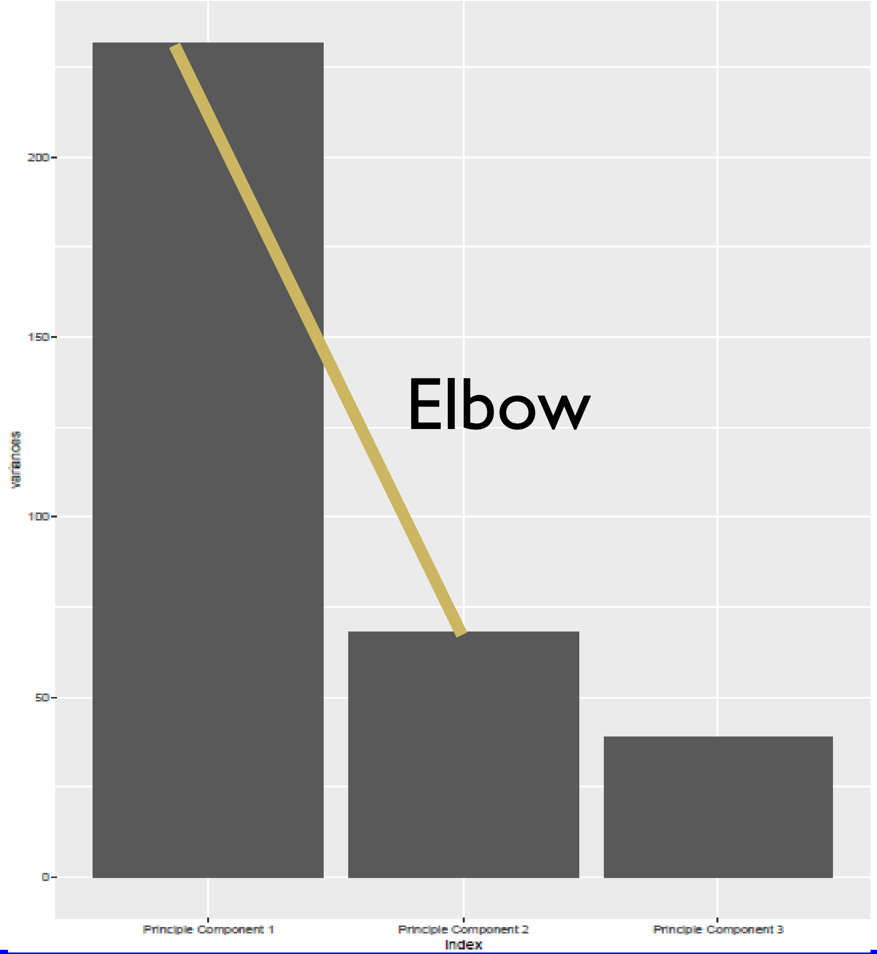
🟧 Low Level

□ High Level

Hierarchical Clustering (Heatmap)



- ✓ Heatmap is a graphical representation of the gene expression profiles data of cervical that illustrated as colors range in a map
- ✓ By calculating the pairwise distance between all of the data
- ✓ Orange: Low level, White: High level

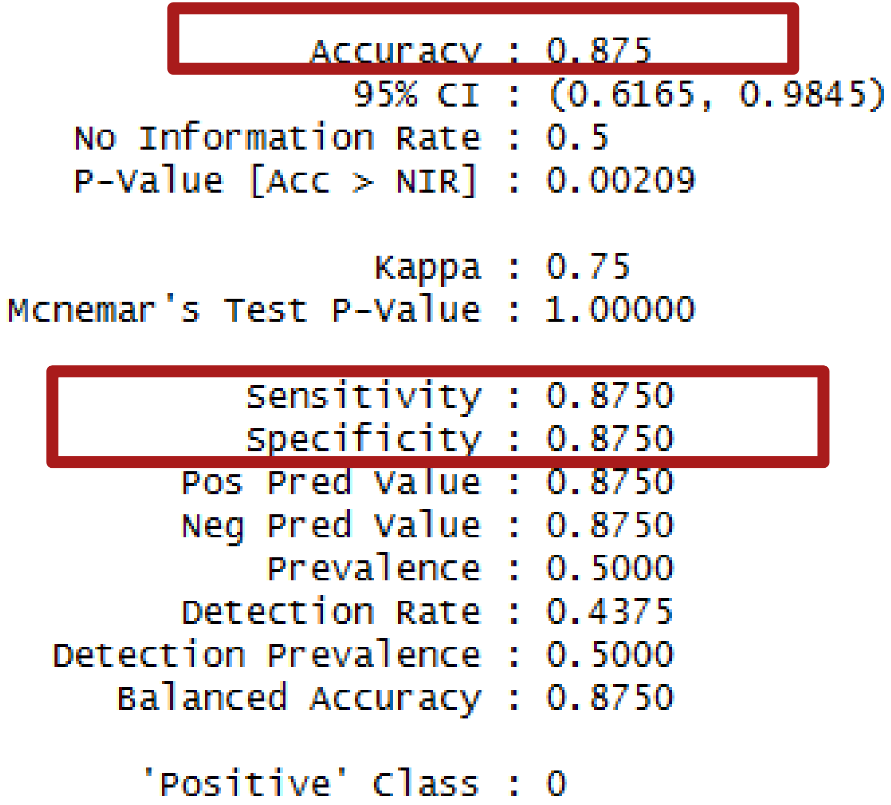


Principle Component Analysis (PCA)

- ✓ Scree plot is a histogram that shows eigenvalues of each principal components (PC)
- ✓ Determine number of PC needed to summaries the dataset
- ✓ Based on scree plot, value of variances decreases dramatically after the first principal components (elbow)
- ✓ Hence, only one PC is sufficient to summarize the dataset

Results (SVM)

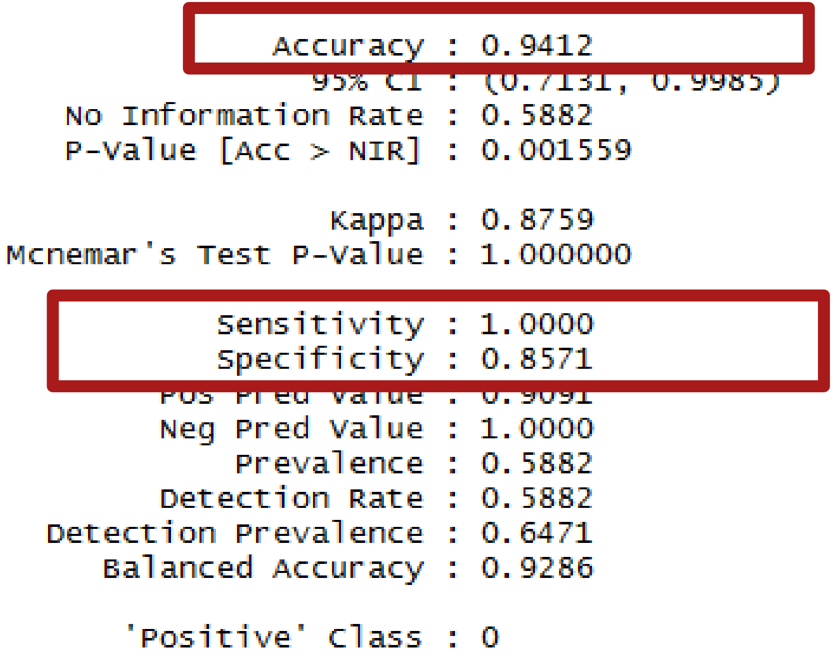
		Reference	
		Normal (0)	Tumor (1)
Prediction	Normal (0)	7	1
	Tumor (1)	1	7



```
               Accuracy : 0.875
                 95% CI : (0.6165, 0.9845)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.00209

                  Kappa : 0.75
 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.8750
            Specificity : 0.8750
         Pos Pred Value : 0.8750
         Neg Pred Value : 0.8750
             Prevalence : 0.5000
         Detection Rate : 0.4375
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.8750

       'Positive' Class : 0
```

Results (RF)

		Reference	
		Normal (0)	Tumor (1)
Prediction	Normal (0)	10	1
	Tumor (1)	0	6

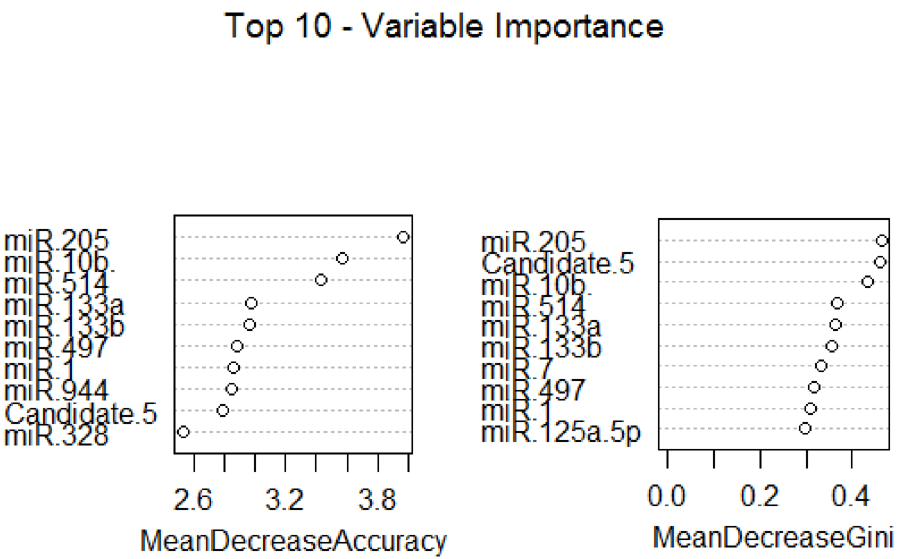


```
               Accuracy : 0.9412
                 95% CI : (0.7131, 0.9985)
    No Information Rate : 0.5882
    P-Value [Acc > NIR] : 0.001559

                  Kappa : 0.8759
 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 1.0000
            Specificity : 0.8571
         Pos Pred Value : 0.9091
         Neg Pred Value : 1.0000
             Prevalence : 0.5882
         Detection Rate : 0.5882
   Detection Prevalence : 0.6471
      Balanced Accuracy : 0.9286

       'Positive' Class : 0
```

Comparison Between SVM and RF

	Support Vector Machine (SVM)	Random Forests (RF)
Accuracy	87.5%	94.12%
Kappa value	0.75	0.8759
Sensitivity	0.8750	1.0000
Specificity	0.8750	0.8571

Variables Importance



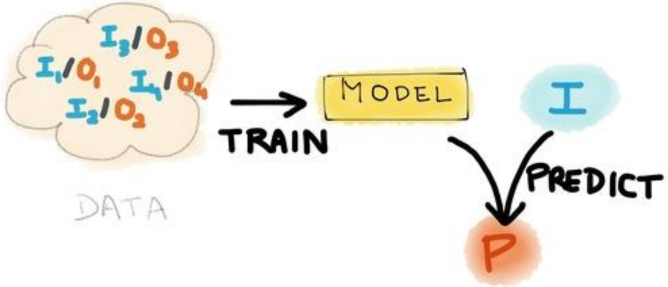
Variables Importance	Mean Decrease Accuracy (MDA)	Variables Importance	Mean Decrease Gini (MDG)
miR. 205	3.96919433	miR. 205	0.462774856
miR. 10b.	3.57406978	Candidate. 5	0.462649239
miR. 514	3.43242093	miR. 10b.	0.434711805
miR 133a	2.98116417	miR. 514	0.367938989
miR. 133b	2.96947145	miR.133a	0.362193567
miR. 497	2.88439595	miR. 133b	0.354346809
miR. 1	2.85682307	miR. 7	0.333107495
miR. 944	2.84431352	miR. 497	0.315540508
	2.7948299		

miR. 205 → Has a role in both normal development and cancer*

*Yue, Z., Yun-shan, Z., & Feng-xia, X. (2016). miR-205 mediates the inhibition of cervical cancer cell proliferation using olmesartan. Journal of the Renin-Angiotensin-Aldosterone System, 17(3), 1470320316663327.

Conclusions

- Random Forests (RF) machine learning algorithms can be successfully used for predicting cervical cancer based on the gene expression profiling data with the microarray dataset
- Model's accuracy obtained is 94.12% which may be acceptable in many applications
- MicroRNA-205 as a novel biomarker for cervical cancer patients
- Big data & machine learning algorithms could be useful in bioinformatics or any other fields


I_3/O_3
I_1/O_1
I_4/O_4
I_2/O_2
DATA
TRAIN
MODEL
I
PREDICT
P