ANALYZING THE RISK FACTOR OF CERVICAL CANCER BY USING THREE MACHINE LEARNING CLASSIFIERS

Name: Wan Nur Syazwani binti Wan Abu Osman

Matric Number: DI170028

Supervisor: Dr. Salama A. Mostafa

Research Background

- Cervical cancer is amongst cancer often occurs against women who attack the reproductive organs.
- Cervical cancer is the third highest cancer in Malaysia after breast and colon cancer.
- There are several early diagnosis methods for cervical cancer screening
- This project will provide an insight on the three classifiers for classifying the risk factor of cervical cancer.

Research Problem

- Its difficult to analyse risk factor of cervical cancer
- Its difficult to determine which is the best machine learning algorithm to be used in classify for risk factor of cervical cancer in terms of accuracy, precision and recall.

Research Objective

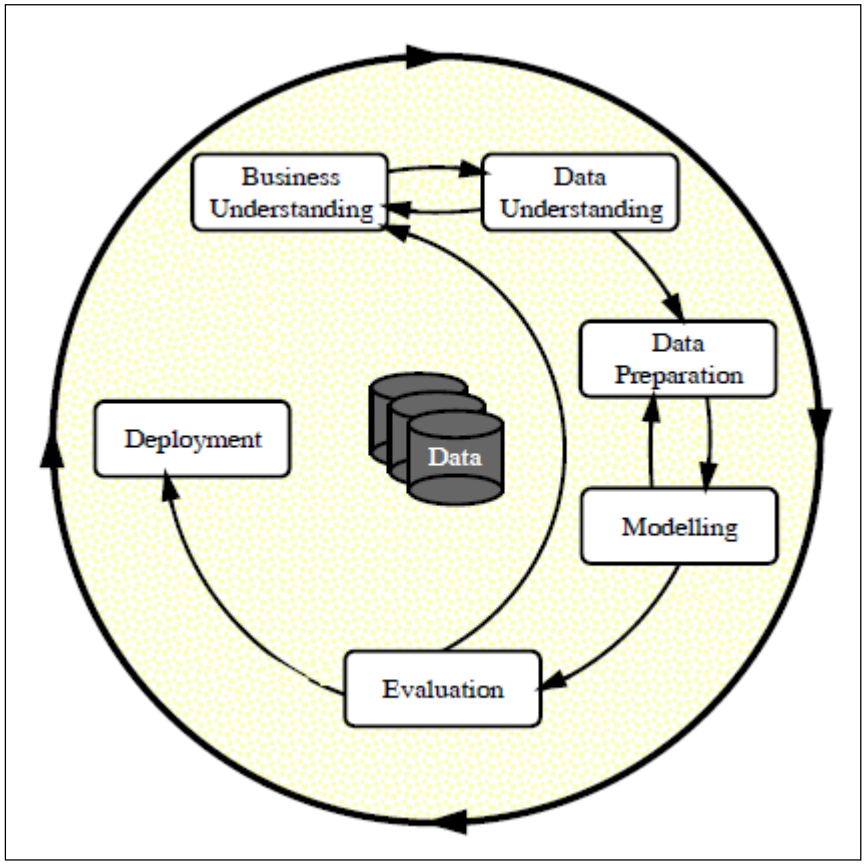
- To investigate the classification methods of cervical cancer in the literature.
- To classify the risk factor of cervical cancer by using three machine learning classifiers
- To evaluate the classification ability of the three machine learning classifiers in terms of accuracy, recall and precision using risk factor cervical cancer dataset

Research Scope

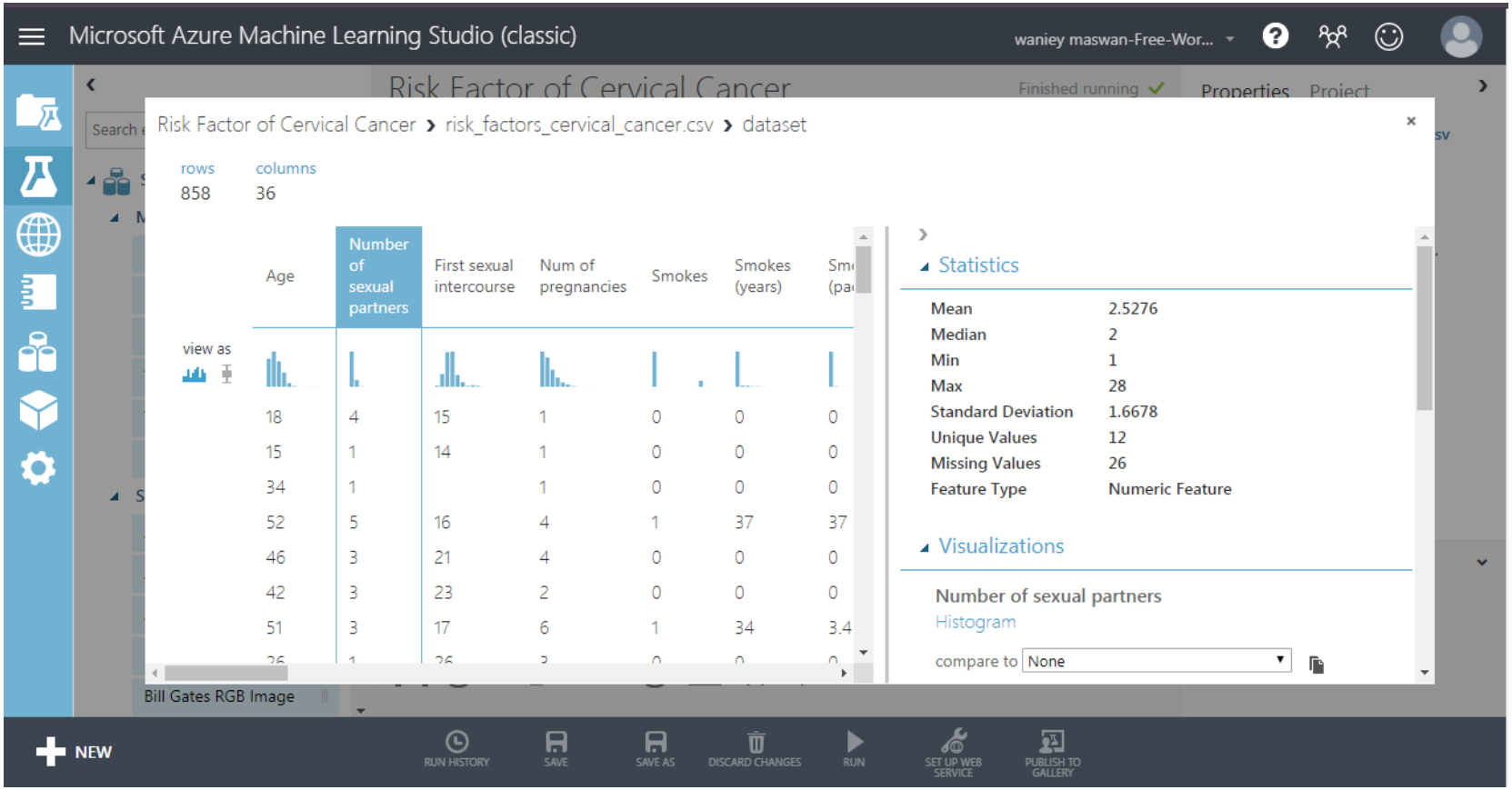
- Data : Risk Factor of Cervical Cancer
- Location : Hospital Universitario de Caracas' in Caracas
- Data Websites : UCI Machine Learning
- Consist : 858 patients, 36 attributes
  : demographic information
  : habits
  : historical medical records
- Method/ : Support Vector Machine Learning (SVM)
  Algorithm : Random Forest (RF)
  : Gradient Boosting Machine (GBM)

Research Methodology

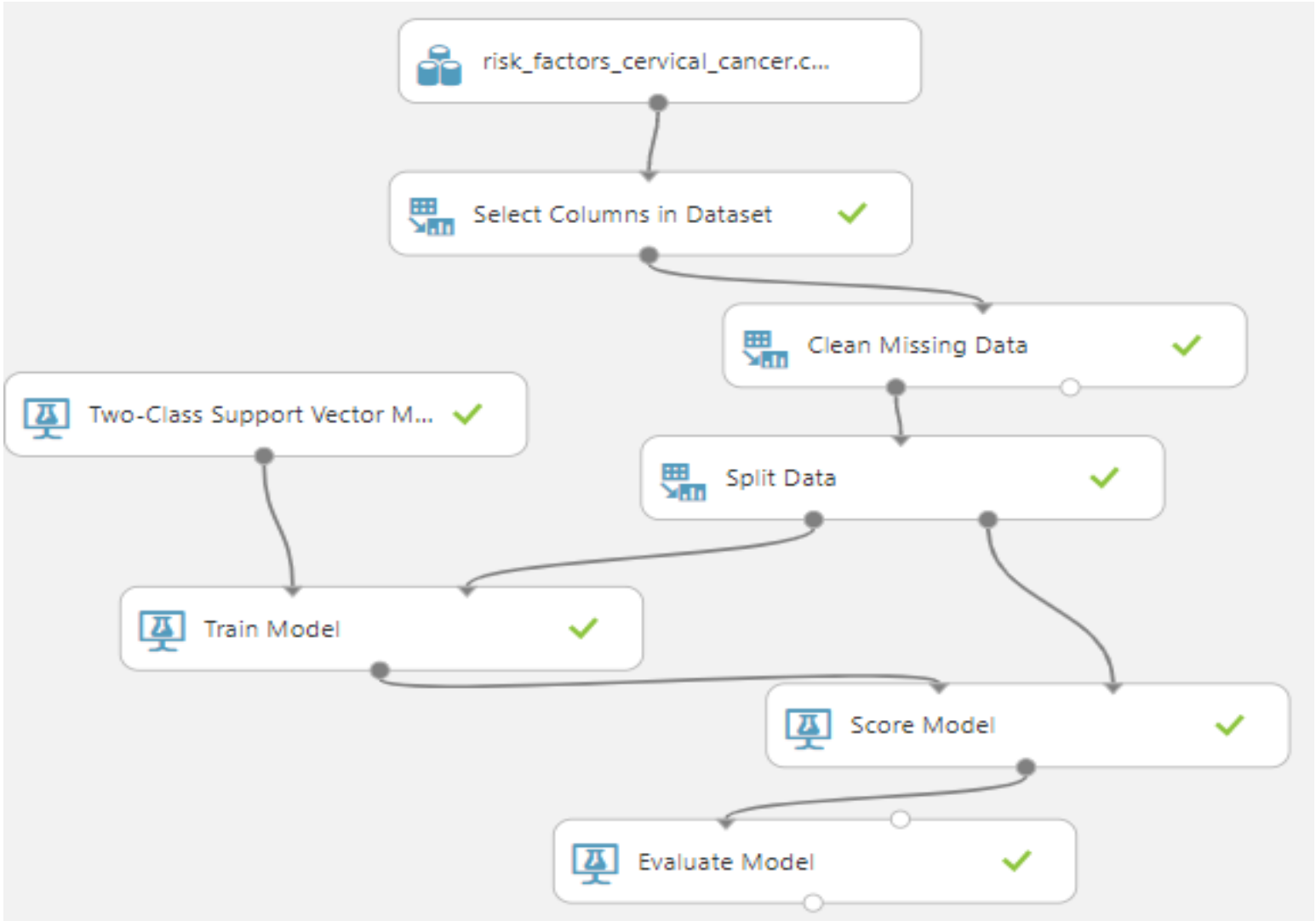
- Methodology -> CRISP-DM
- methodology that provides a structured approach to planning a data mining project



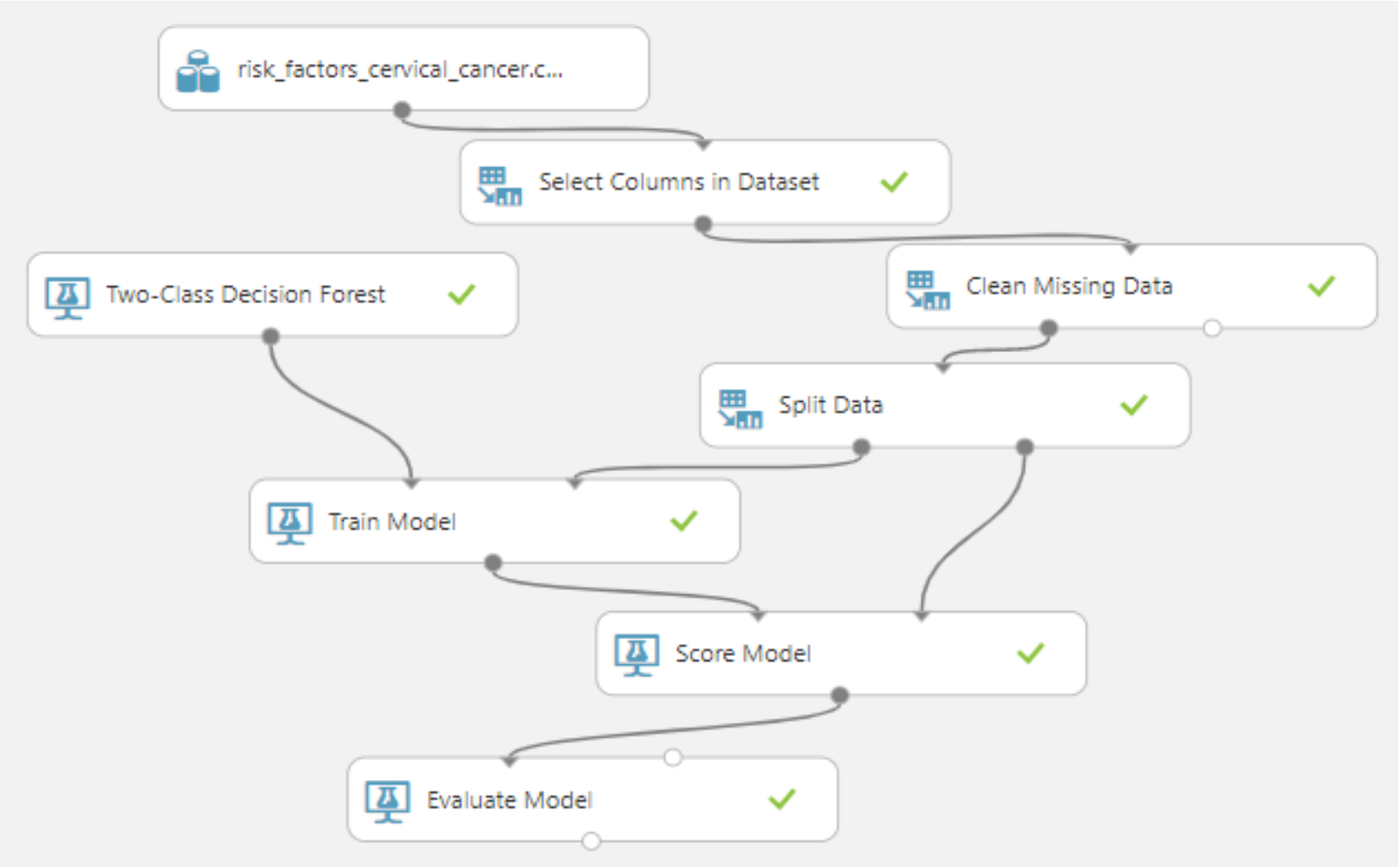
Data set visualization in MS Azure



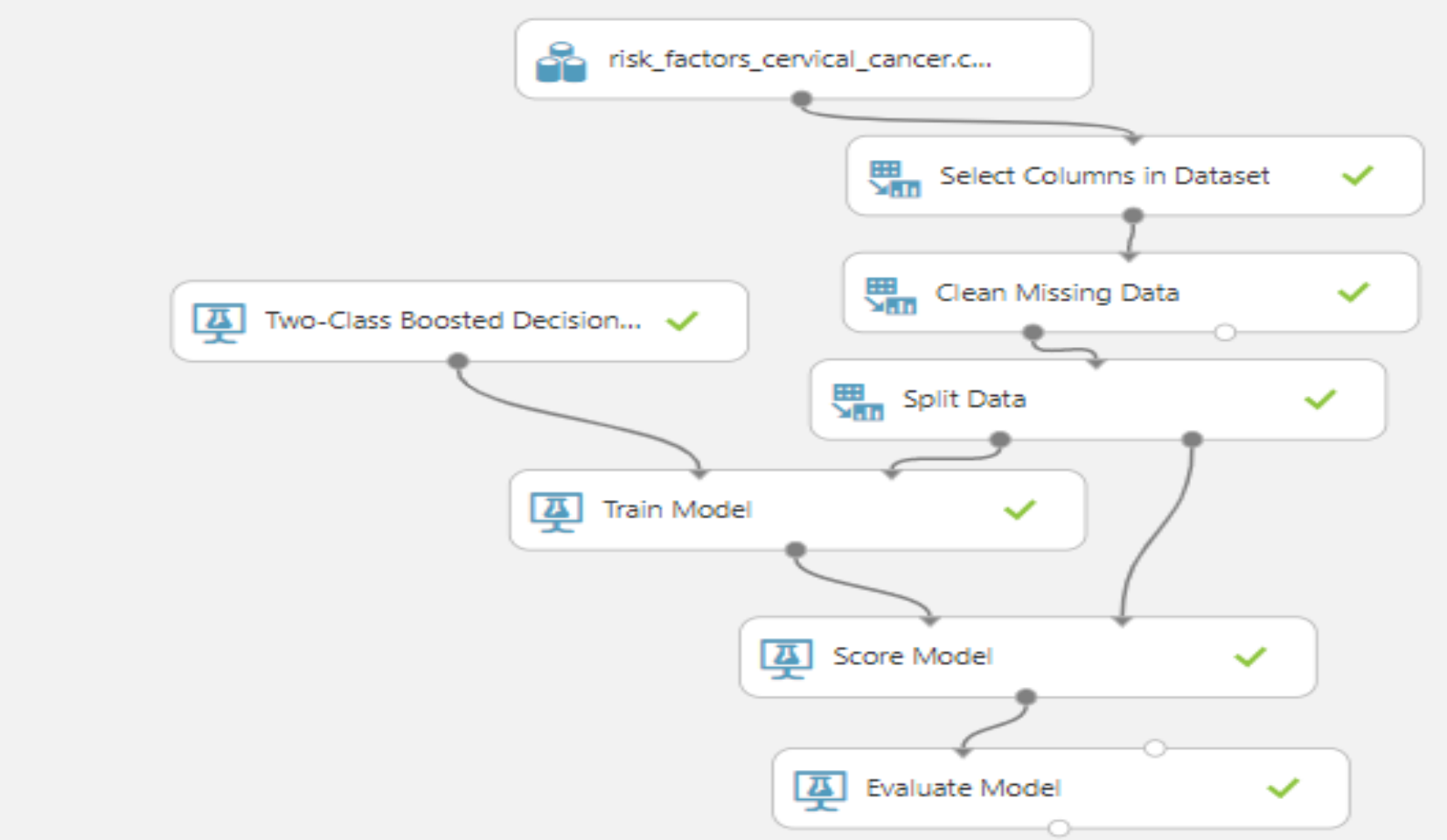
Build Model Support Vector Machine



Build Model Random Forest



Build Model Gradient Boosting Machine

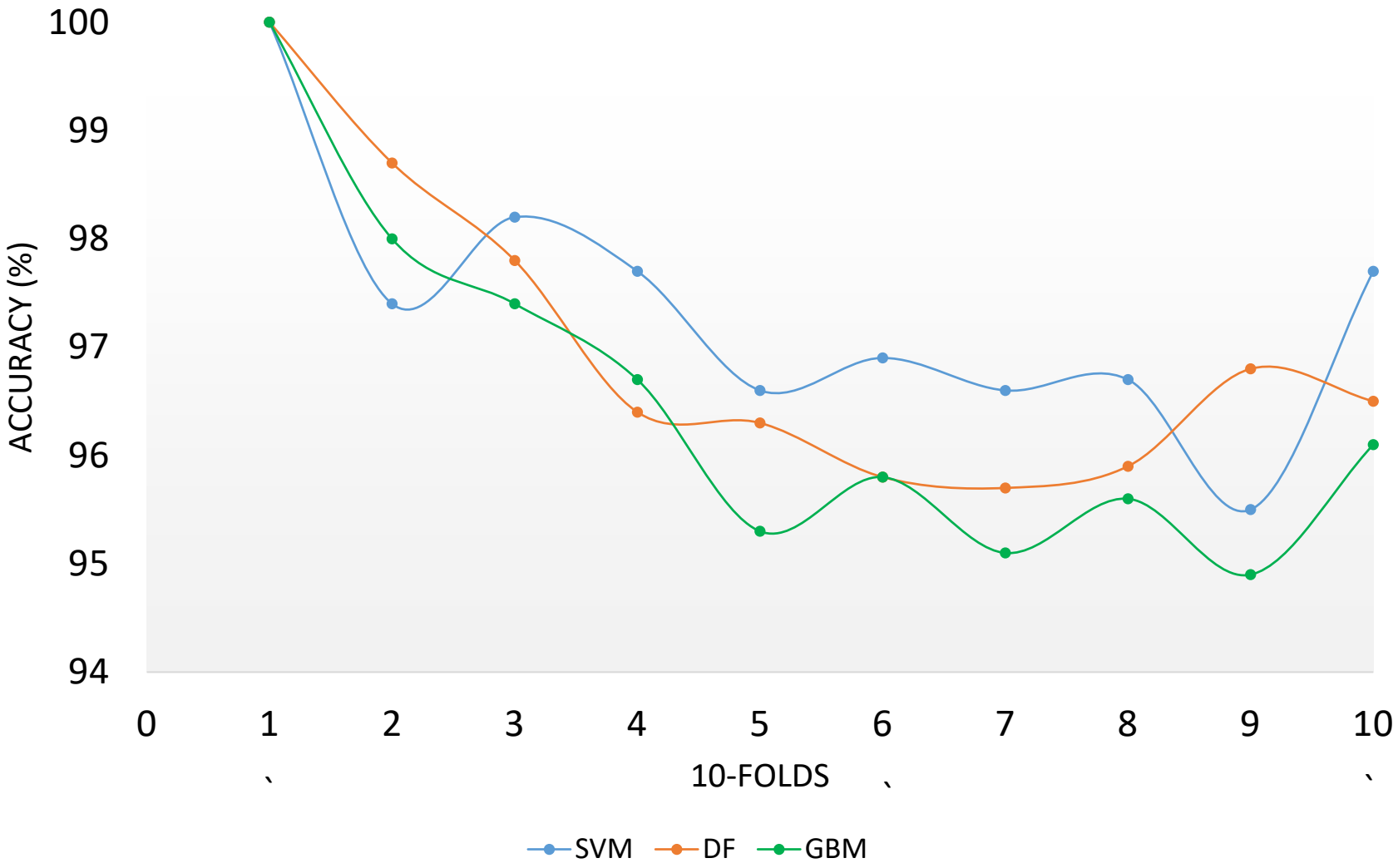


Experimental Result

Result of Accuracy

Test	Split data	SVM	DF	GBM
1	90:10	100	100	100
2	80:20	97.4	98.7	98.0
3	70:30	98.2	97.8	97.4
4	60:40	97.7	96.4	96.7
5	50:50	96.6	96.3	95.3
6	40:60	96.9	95.8	95.8
7	30:70	96.6	95.7	95.1
8	20:80	96.7	95.9	95.6
9	10:90	95.5	96.8	94.9
10	66:34	97.7	96.5	96.1
Average		97.33	96.99	96.49
Standard Deviation		1.2093	1.4146	1.5906

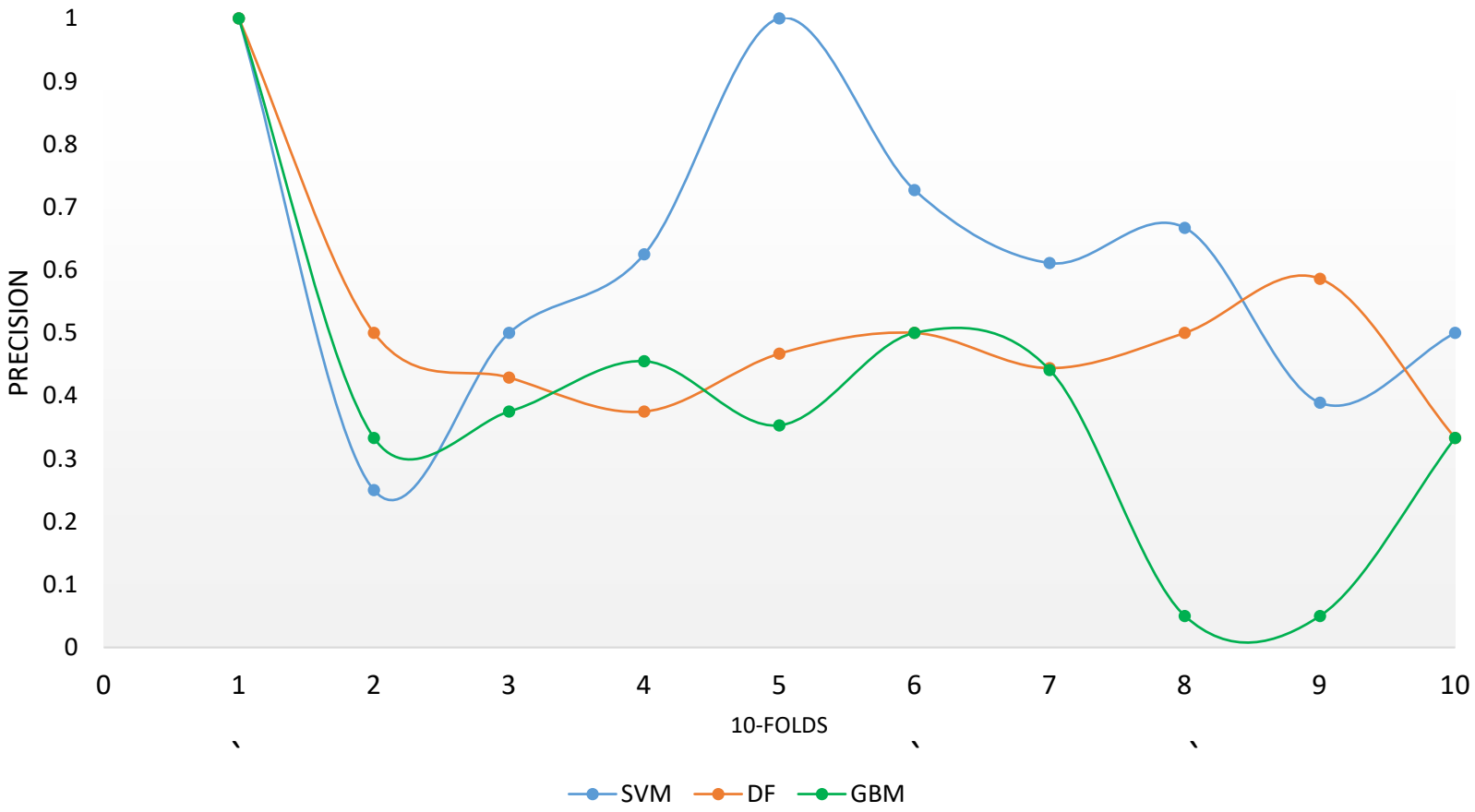
Graph for the Accuracy Result



Result of Precision

Test	Split data	SVM	DF	GBM
1	90:10	1.000	1.000	1.000
2	80:20	0.250	0.500	0.333
3	70:30	0.500	0.429	0.375
4	60:40	0.625	0.375	0.455
5	50:50	1.000	0.467	0.353
6	40:60	0.727	0.500	0.500
7	30:70	0.611	0.444	0.441
8	20:80	0.667	0.500	0.050
9	10:90	0.389	0.586	0.050
10	66:34	0.500	0.333	0.333
Average		0.627	0.513	0.389
Standard Deviation		0.241	0.185	0.265

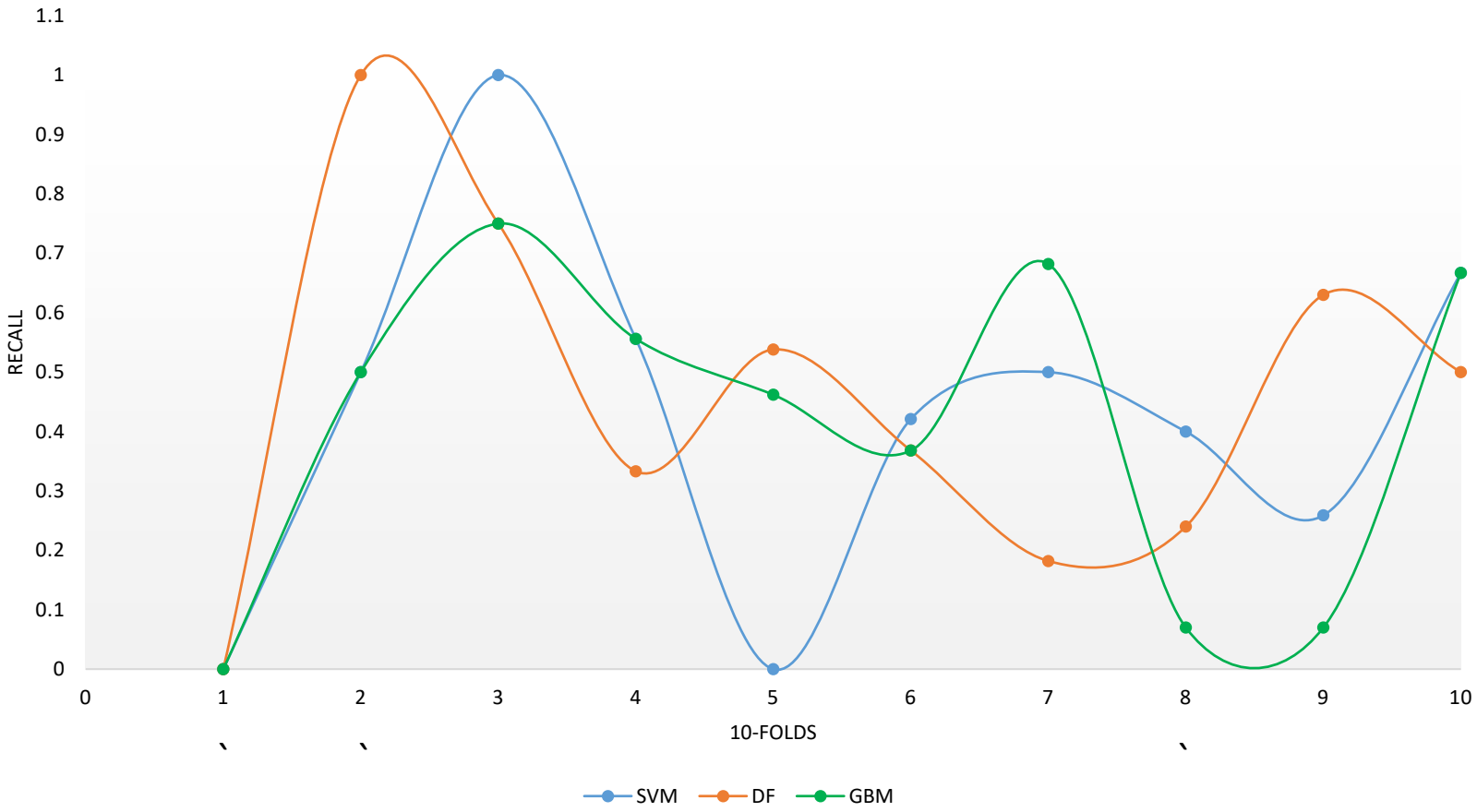
Graph for the Precision Result



Result of Recall

Test	Split data	SVM	DF	GBM
1	90:10	0.000	0.000	0.000
2	80:20	0.500	1.000	0.500
3	70:30	1.000	0.750	0.750
4	60:40	0.556	0.333	0.556
5	50:50	0.000	0.538	0.462
6	40:60	0.421	0.368	0.368
7	30:70	0.500	0.182	0.682
8	20:80	0.400	0.240	0.070
9	10:90	0.259	0.630	0.070
10	66:34	0.667	0.500	0.667
Average		0.430	0.454	0.413
Standard Deviation		0.284	0.278	0.262

Graph for the Recall Result



Conclusion

- This research is present about analysis on risk factor of cervical cancer data using classification algorithms
- The evaluation metric such as accuracy, precision and recall are measured for the given dataset to estimate the performance of each classification techniques
- As a result, SVM was the highest rate of accuracy compare to DF and GBM. The results show that SVM (97.33%) is slightly better compared to DF (96.99%) and GBM (96.49%) in term of accuracy.

Thank You