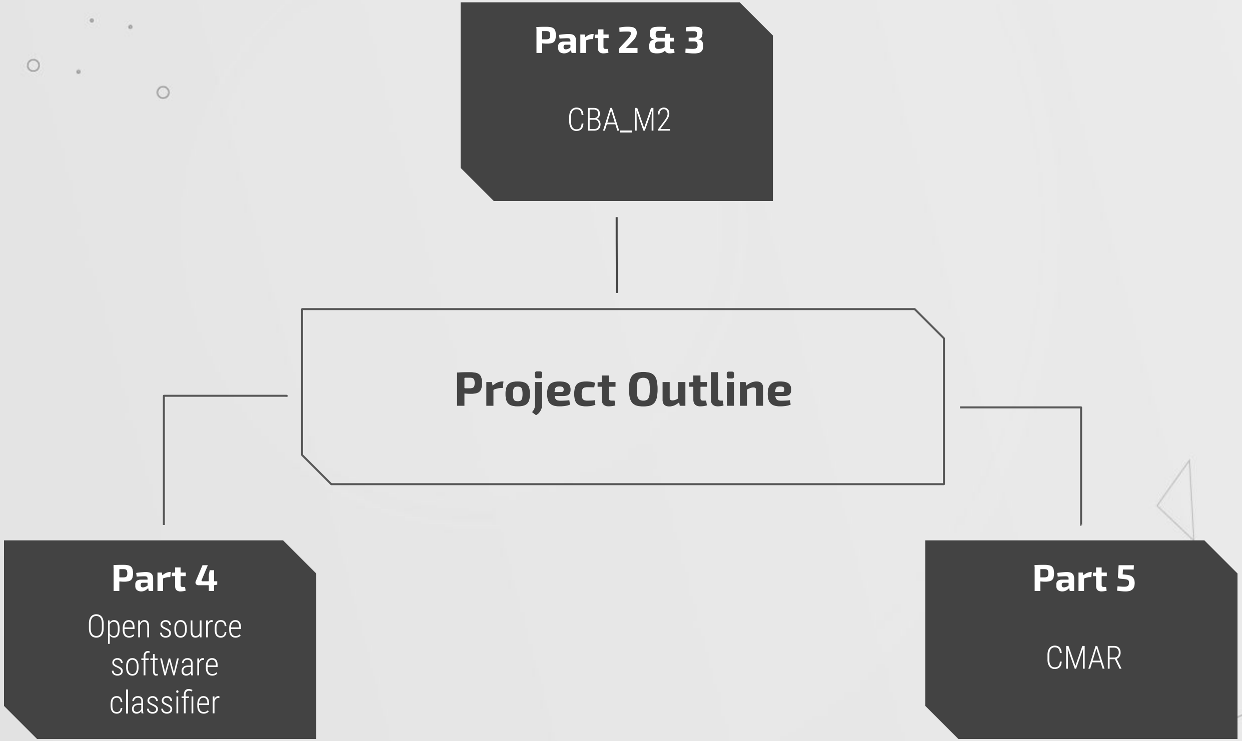
CZ4032
Data Analytics & Mining
Project I, Group 24

Chang Heen Sunn, Cao Shuwen, Huang Runtao, Yin Jia Rui

01

INTRODUCTION



Project Outline

Part 2 & 3

CBA_M2

Part 4

Open source software classifier

Part 5

CMAR

DATASET

- Glass
- Wine
- Iris
- Pima
- Tic-tac-toe

Dataset in paper

- Caesarian
- Car

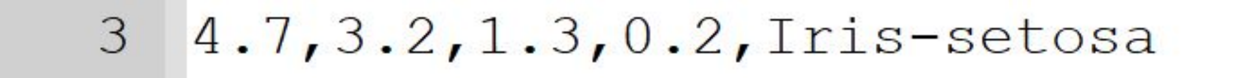
Other dataset

02

IMPLEMENTATION OF CBA_M2

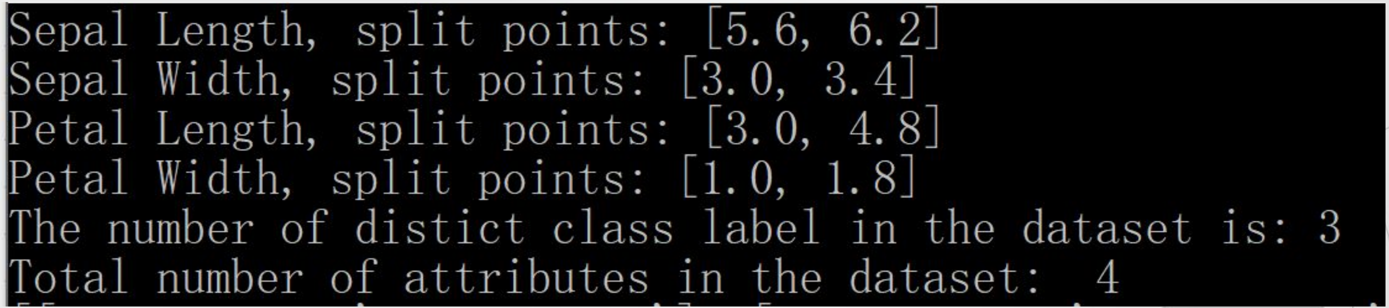
Data Preprocessing (e.g. Iris data)

4 attributes & 1 class label



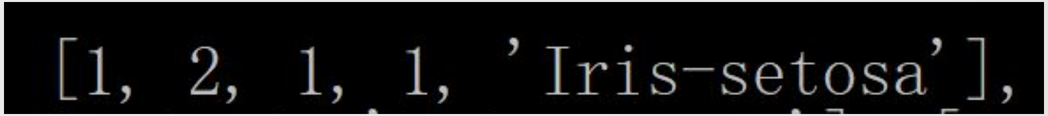
```
4.7,3.2,1.3,0.2,Iris-setosa
```

Split points:



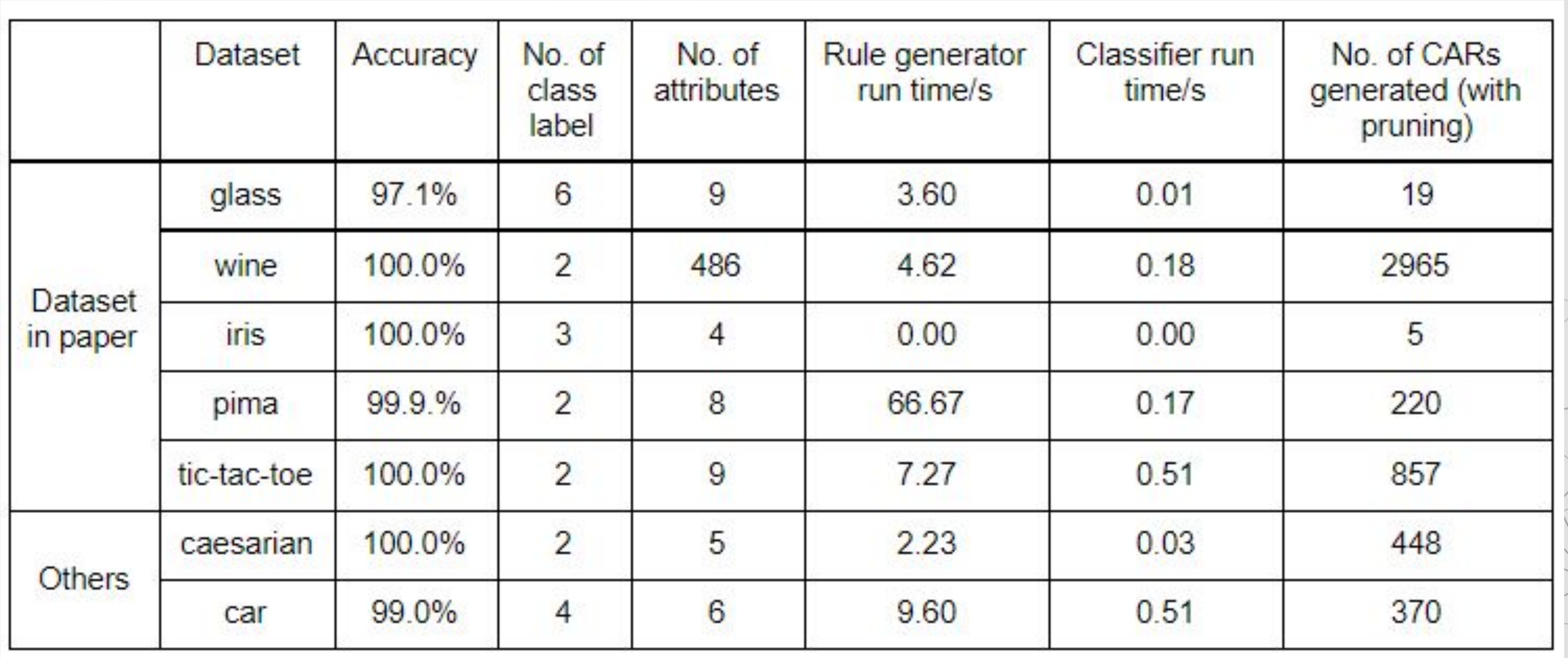
```
Sepal Length, split points: [5.6, 6.2]
Sepal Width, split points: [3.0, 3.4]
Petal Length, split points: [3.0, 4.8]
Petal Width, split points: [1.0, 1.8]
The number of distict class label in the dataset is: 3
Total number of attributes in the dataset:  4
```

Output:



```
 [1, 2, 1, 1, 'Iris-setosa'],
```

Results of CBA_M2 algorithm



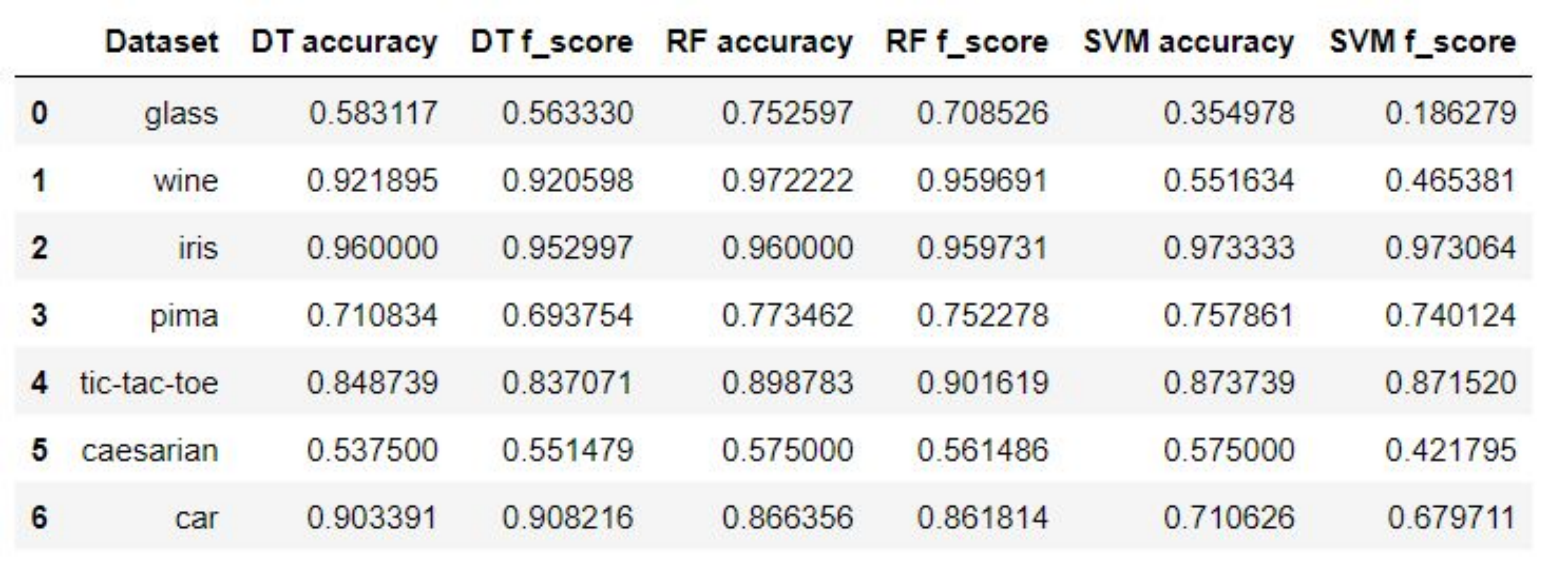
	Dataset	Accuracy	No. of class label	No. of attributes	Rule generator run time/s	Classifier run time/s	No. of CARs generated (with pruning)
Dataset in paper	glass	97.1%	6	9	3.60	0.01	19
	wine	100.0%	2	486	4.62	0.18	2965
	iris	100.0%	3	4	0.00	0.00	5
	pima	99.9.%	2	8	66.67	0.17	220
	tic-tac-toe	100.0%	2	9	7.27	0.51	857
Others	caesarian	100.0%	2	5	2.23	0.03	448
	car	99.0%	4	6	9.60	0.51	370

03

OPEN SOFTWARE CLASSIFIER

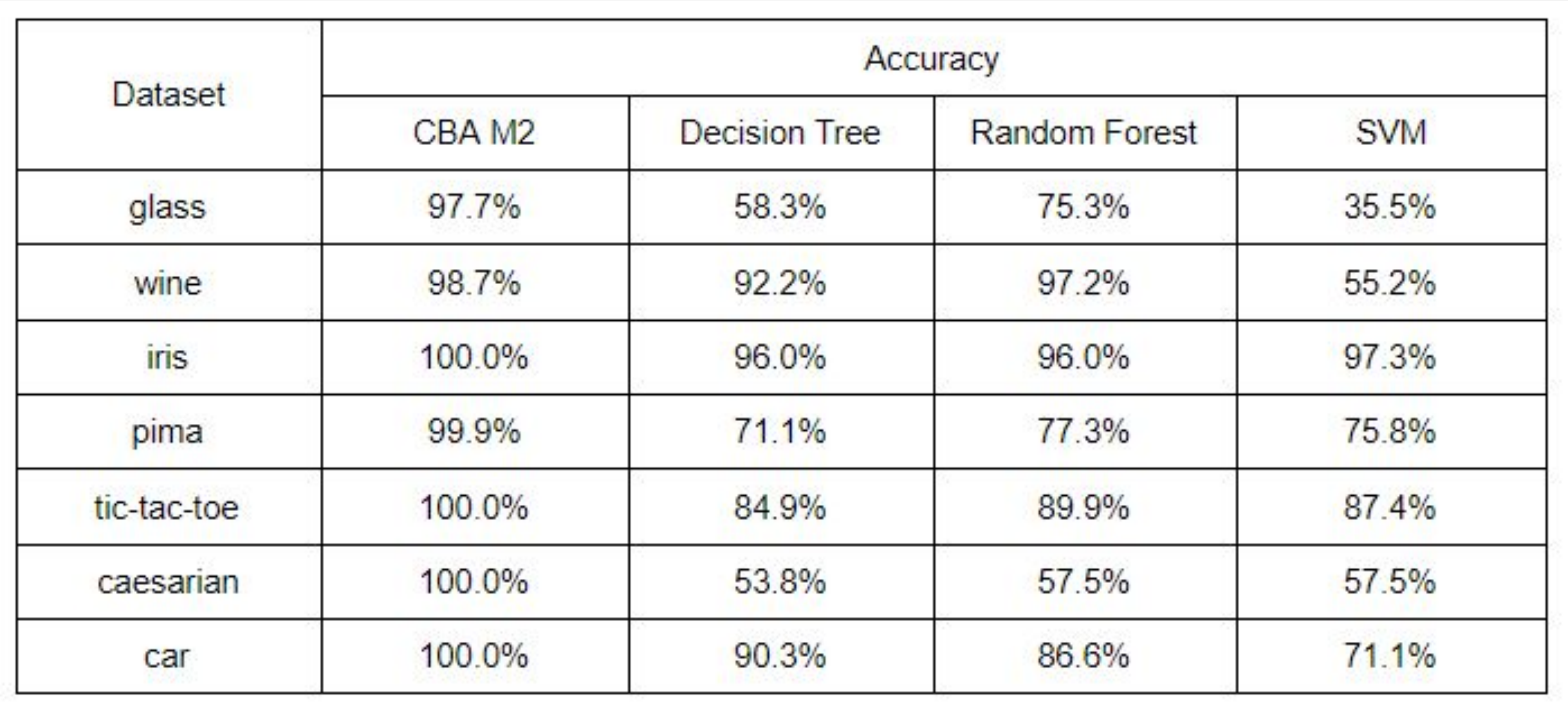
Decision Tree, Random Forest &
Support Vector Machine

Results of
Open source software Classifiers



	Dataset	DT accuracy	DT f_score	RF accuracy	RF f_score	SVM accuracy	SVM f_score
0	glass	0.583117	0.563330	0.752597	0.708526	0.354978	0.186279
1	wine	0.921895	0.920598	0.972222	0.959691	0.551634	0.465381
2	iris	0.960000	0.952997	0.960000	0.959731	0.973333	0.973064
3	pima	0.710834	0.693754	0.773462	0.752278	0.757861	0.740124
4	tic-tac-toe	0.848739	0.837071	0.898783	0.901619	0.873739	0.871520
5	caesarian	0.537500	0.551479	0.575000	0.561486	0.575000	0.421795
6	car	0.903391	0.908216	0.866356	0.861814	0.710626	0.679711

Comparison of results
CBA M2 vs Open source classifiers



Dataset	Accuracy			
	CBA M2	Decision Tree	Random Forest	SVM
glass	97.7%	58.3%	75.3%	35.5%
wine	98.7%	92.2%	97.2%	55.2%
iris	100.0%	96.0%	96.0%	97.3%
pima	99.9%	71.1%	77.3%	75.8%
tic-tac-toe	100.0%	84.9%	89.9%	87.4%
caesarian	100.0%	53.8%	57.5%	57.5%
car	100.0%	90.3%	86.6%	71.1%

04

ADVANCED ALGORITHM

Classification based on
Multiple Association Rules Method (CMAR)

Implemented Algorithms in CMAR

- Rule Mining: FP-tree with FP-growth

- Pruning:

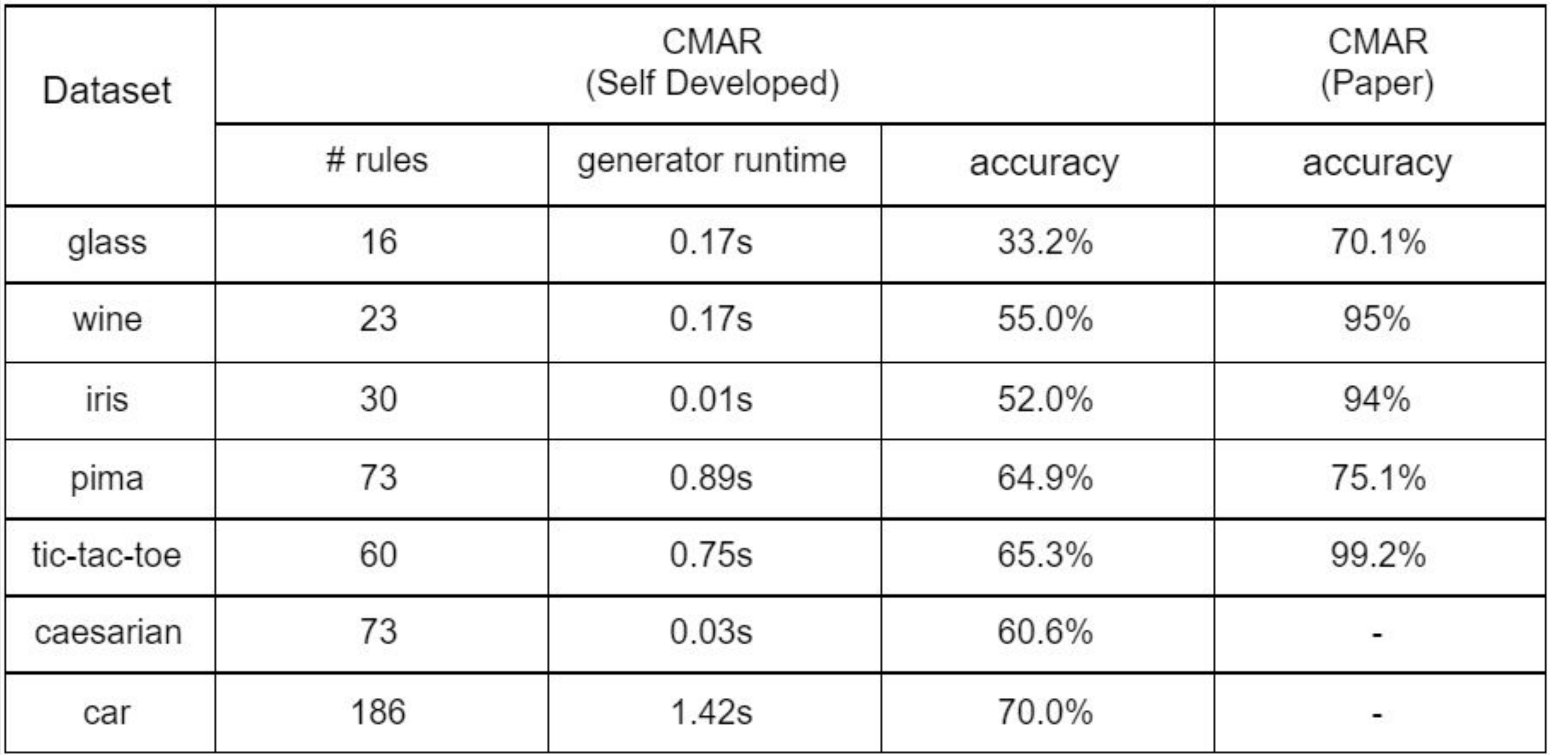
  1. Prune more specific and low confidence rules

  2. Prune rules with X^2 lower than 0.05 probability threshold

  3. Keep the rules with higher rank which can cover the database a few times

- Classifier: Compare weighted X^2

Results of CMAR Algorithm



Dataset	CMAR (Self Developed)			CMAR (Paper)
	# rules	generator runtime	accuracy	accuracy
glass	16	0.17s	33.2%	70.1%
wine	23	0.17s	55.0%	95%
iris	30	0.01s	52.0%	94%
pima	73	0.89s	64.9%	75.1%
tic-tac-toe	60	0.75s	65.3%	99.2%
caesarian	73	0.03s	60.6%	-
car	186	1.42s	70.0%	-

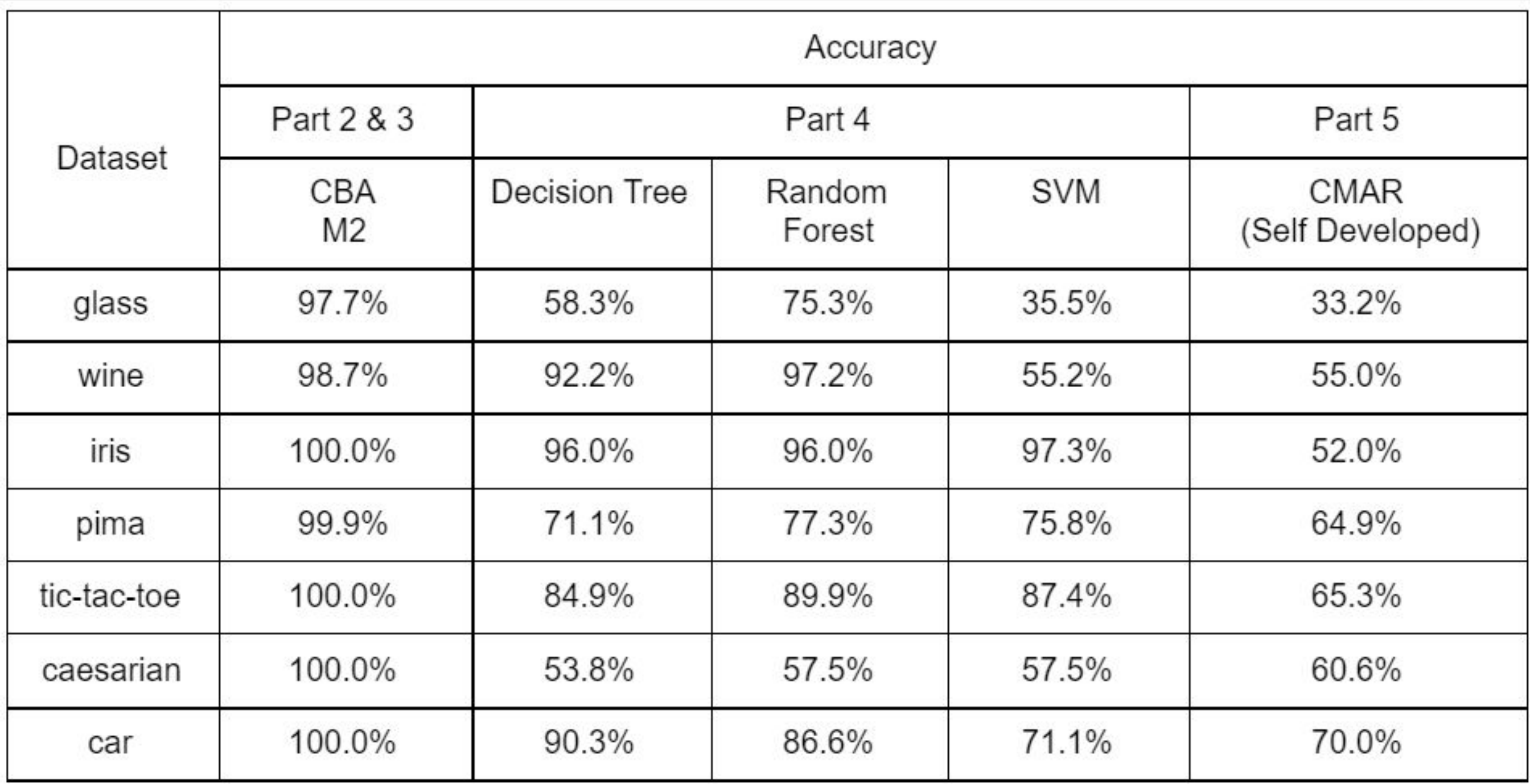
CMAR Results Discussion

Low accuracy: may be resulted from the rule mining part

Fast rule generator: use of compact structures (FP-tree & CR-tree)

05

CONCLUSION



Dataset	Accuracy				
	Part 2 & 3	Part 4			Part 5
	CBA M2	Decision Tree	Random Forest	SVM	CMAR (Self Developed)
glass	97.7%	58.3%	75.3%	35.5%	33.2%
wine	98.7%	92.2%	97.2%	55.2%	55.0%
iris	100.0%	96.0%	96.0%	97.3%	52.0%
pima	99.9%	71.1%	77.3%	75.8%	64.9%
tic-tac-toe	100.0%	84.9%	89.9%	87.4%	65.3%
caesarian	100.0%	53.8%	57.5%	57.5%	60.6%
car	100.0%	90.3%	86.6%	71.1%	70.0%

- Best performing: CBA M2
- Worst performing: CMAR

Takeaways

Implemented classification methods: CBA M2, CMAR, Decision Tree, Random Forest, SVM

Gained a greater understanding of FP-growth and Apriori Algorithms

06

CONTRIBUTION & REFERENCE

CONTRIBUTION

Name	Assigned Tasks
Cao Shuwen	Data preprocessing & CBA classifier
Chang Heen Sunn	CBA classifier & open source software classifier
Huang Runtao	CBA Rule Generator & CMAR algorithm

P/S: In fact, we have another member in our group (Yin Jia Rui) but she never contribute anything to the project.

References

1. B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In KDD’98, New York, NY, Aug. 1998.

2. Wenmin Li, Jiawei Han and Jian Pei, "CMAR: accurate and efficient classification based on multiple class-association rules," Proceedings 2001 IEEE International Conference on Data Mining, 2001, pp. 369-376, doi: 10.1109/ICDM.2001.989541.

3. T. D. V. Swinscow, “Statistics at square one: The BMJ,” *The BMJ | The BMJ: leading general medical journal. Research. Education. Comment*, 28-Oct-2020. [Online]. Available: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one. [Accessed: 20-Oct-2021].

4. https://github.com/Williano/Data-Mining/tree/b24247ff3cb8eb0227885dd27287d4dace7aa629/wine_data_mining_research/association_classification

THANKS

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.