1 The Data

- Two `Source` datasets TS3-Sapphire.arff and TS6-Sapphire.arff (This is for strain *Plasmodium vivax*)
- The TL algorithm supports having multiple source datasets as it assigns weights to instances as well as tasks
- I used data from the Venus Channel (Venus Active/Inactive) for assay TS6 to create `Target` and `Test` datasets
- This dataset has 1435 instances and the percentage of `Active` instances is ˜2% (27 instances)
- I started with a very small `Target` dataset of all `Inactives` and doubled the size at each iteration
- As the size increased, I started to include Active instances (with the same percentage)
- The `Target` dataset was used as `Training` dataset to build TL, NB, J48, SMO and KNN models
- Details of the various datasets are shown in the following table:

Setting No.	**Target Dataset (Training)**	**Test Dataset**
1	Size=3 (3 Inactive + 0 Active)	Size=1432 (1405 Inactive + 27 Active)
2	Size=6 (6 Inactive + 0 Active)	Size=1429 (1402 Inactive + 27 Active)
3	Size=12 (12 Inactive + 0 Active)	Size=1423 (1396 Inactive + 27 Active)
4	Size=24 (24 Inactive + 0 Active)	Size=1411 (1384 Inactive + 27 Active)
5	Size=49 (48 Inactive + 1 Active)	Size=1386 (1360 Inactive + 26 Active)
6	Size=98 (96 Inactive + 2 Active)	Size=1337 (1312 Inactive + 25 Active)
7	Size=196 (192 Inactive + 4 Active)	Size=1239 (1216 Inactive + 23 Active)
8	Size=392 (384 Inactive + 8 Active)	Size=1043 (1024 Inactive + 19 Active)

1.1 Experimental Stat Results for Setting Number 1:

- Target (Training) Dataset: Size=3 (3 Inactive + 0 Active)
- Testing Dataset: Size=1432 (1405 Inactive + 27 Active)
- In this experiment I did 3 fold cross validation

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	100	0	?	1	0	0	0	0	0	0	0	0
NB	100	0	?	1	0	0	0	0	0	0	0	0
J48	100	0	?	1	0	0	0	0	0	0	0	0
SMO	100	0	?	1	0	0	0	0	0	0	0	0
IBk	100	0	?	1	0.16	0.16	66.66	66.66	0	0	0	0

But when evaluating using the test set, the results were as follows:

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	99.3	0.69	0.99	0.81	0	0.08	3.3	36.88	0.81	0.81	0.81	0
NB	98.11	1.88	0.5	0	0.01	0.13	8.92	60.61	0	0	0	0.01
J48	98.11	1.88	0.5	0	0.01	0.13	8.92	60.61	0	0	0	0.01
SMO	98.11	1.88	0.5	0	0.01	0.13	8.92	60.61	0	0	0	0.01
IBk	98.11	1.88	0.5	0	0.1	0.15	50.32	67.94	0	0	0	0.01

Conclusion: TL is a clear winner

1.2 Experimental Stat Results for Setting Number 2:

- Target (Training) Dataset:Size=6 (6 Inactive + 0 Active)
- Testing Dataset: Size=1429 (1402 Inactive + 27 Active)
- In this experiment I did 3 fold cross validation

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	100	0	?	1	0	0	0	0	0	0	0	0
NB	100	0	?	1	0	0	0	0	0	0	0	0
J48	100	0	?	1	0	0	0	0	0	0	0	0
SMO	100	0	?	1	0	0	0	0	0	0	0	0
IBk	100	0	?	1	0.05	0.05	33.33	33.33	0	0	0	0

But when evaluating using the test set, the results were as follows:

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	99.3	0.69	0.99	0.81	0	0.08	5.03	48.46	0.81	0.81	0.81	0
NB	98.11	1.88	0.5	0	0.01	0.13	13.57	79.63	0	0	0	0.01
J48	98.11	1.88	0.5	0	0.01	0.13	13.57	79.63	0	0	0	0.01
SMO	98.11	1.88	0.5	0	0.01	0.13	13.57	79.63	0	0	0	0.01
IBk	98.11	1.88	0.5	0	0.04	0.13	35.18	79.2	0	0	0	0.01

Conclusion: TL is a clear winner

1.3 Experimental Stat Results for Setting Number 3:

- Target (Training) Dataset: Size=12 (12 Inactive + 0 Active)
- Testing Dataset: Size=1423 (1396 Inactive + 27 Active)
- In this experiment I did 10 fold cross validation

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	100	0	?	1	0	0	0	0	0	0	0	0
NB	100	0	?	1	0	0	0	0	0	0	0	0
J48	100	0	?	1	0	0	0	0	0	0	0	0
SMO	100	0	?	1	0	0	0	0	0	0	0	0
IBk	100	0	?	1	0.01	0.01	22.9	22.9	0	0	0	0

But when evaluating using the test set, the results were as follows:

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	99.29	0.7	0.99	0.81	0	0.08	8.01	57.34	0.81	0.81	0.81	0
NB	98.1	1.89	0.5	0	0.01	0.13	21.63	94.23	0	0	0	0.01
J48	98.1	1.89	0.5	0	0.01	0.13	21.63	94.23	0	0	0	0.01
SMO	98.1	1.89	0.5	0	0.01	0.13	21.63	94.23	0	0	0	0.01
IBk	98.1	1.89	0.5	0	0.03	0.13	39.33	93.35	0	0	0	0.01

Conclusion: TL is a clear winner

1.4 Experimental Stat Results for Setting Number 4:

- Target (Training) Dataset: Size=24 (24 Inactive + 0 Active)
- Testing Dataset: Size=1411 (1384 Inactive + 27 Active)
- In this experiment I did 10 fold cross validation

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	95.83	4.16	?	0	0.04	0.2	97.87	479.36	0	0	0	0.04
NB	100	0	?	1	0	0	0	0	0	0	0	0
J48	100	0	?	1	0	0	0	0	0	0	0	0
SMO	100	0	?	1	0	0	0	0	0	0	0	0
IBk	100	0	?	1	0	0	21.46	21.46	0	0	0	0

But when evaluating using the test set, the results were as follows:

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	99.36	0.63	0.99	0.82	0	0.07	11.37	57.16	0.84	0.81	0.83	0
NB	98.08	1.91	0.5	0	0.01	0.13	34.09	99.98	0	0	0	0.01
J48	98.08	1.91	0.5	0	0.01	0.13	34.09	99.98	0	0	0	0.01
SMO	98.08	1.91	0.5	0	0.01	0.13	34.09	99.98	0	0	0	0.01
IBk	98.08	1.91	0.5	0	0.02	0.13	48.13	99.33	0	0	0	0.01

Conclusion: TL is a clear winner

1.5 Experimental Stat Results for Setting Number 5:

- Target (Training) Dataset: Size=49 (48 Inactive + 1 Active)
- Testing Dataset: Size=1386 (1360 Inactive + 26 Active)
- In this experiment I did 10 fold cross validation

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	95.91	4.08	0.16	-0.03	0.04	0.2	66.89	137.85	0	0	0	0.04
NB	97.95	2.04	0.5	0	0.02	0.14	33.61	97.97	0	0	0	0.02
J48	97.95	2.04	0.04	0	0.04	0.14	67.16	99.07	0	0	0	0.02
SMO	95.91	4.08	0.48	-0.03	0.04	0.2	67.23	138.55	0	0	0	0.04
IBk	97.95	2.04	0.53	0	0.02	0.14	47.38	99.58	0	0	0	0.02

But when evaluating using the test set, the results were as follows:

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	99.35	0.64	0.99	0.82	0	0.07	11.1	55.07	0.81	0.84	0.83	0
NB	98.12	1.87	0.5	0	0.01	0.13	33.19	99.82	0	0	0	0.01
J48	98.12	1.87	0.5	0	0.03	0.13	67.96	98.88	0	0	0	0.01
SMO	99.42	0.57	0.92	0.84	0	0.07	10.21	55.37	0.84	0.84	0.84	0
IBk	98.12	1.87	0.96	0	0.01	0.11	34.77	81.65	0	0	0	0.01

Conclusion: TL and SMO perform better than other models

1.6 Experimental Stat Results for Setting Number 6:

- Target (Training) Dataset: Size=98 (96 Inactive + 2 Active)
- Testing Dataset: Size=1337 (1312 Inactive + 25 Active)
- In this experiment I did 10 fold cross validation

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	98.97	1.02	0.99	0.79	0.01	0.1	20.22	70.58	0.66	1	0.8	0.01
NB	95.91	4.08	0.09	-0.03	0.04	0.2	80.82	141.16	0	0	0	0.04
J48	96.93	3.06	0.89	-0.02	0.03	0.17	64.29	121.37	0	0	0	0.03
SMO	97.95	2.04	0.74	0.48	0.02	0.14	40.41	99.82	0.5	0.5	0.5	0.02
IBk	97.95	2.04	0.58	0	0.02	0.14	48.72	100.61	0	0	0	0.02

But when evaluating using the test set, the results were as follows:

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	99.32	0.67	0.99	0.81	0	0.08	14.68	60.48	0.83	0.8	0.81	0
NB	99.62	0.37	0.98	0.9	0	0.06	7.86	44.98	0.85	0.96	0.9	0
J48	98.27	1.72	0.54	0.14	0.01	0.13	36.15	96.49	1	0.08	0.14	0.01
SMO	99.17	0.82	0.79	0.72	0	0.09	17.29	66.72	0.93	0.6	0.73	0
IBk	98.13	1.86	0.86	0	0.01	0.11	35.27	81.06	0	0	0	0.01

Conclusion: TL and NB perform better than other models

1.7 Experimental Stat Results for Setting Number 7:

- Target (Training) Dataset: Size=196 (192 Inactive + 4 Active)
- Testing Dataset: Size=1239 (1216 Inactive + 23 Active)
- In this experiment I did 10 fold cross validation

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	100	0	1	1	0	0	0.01	0.06	1	1	1	0
NB	97.95	2.04	0.99	0.65	0.02	0.14	45.12	100.61	0.5	1	0.66	0.02
J48	98.97	1.02	0.87	0.74	0.01	0.1	22.56	71.14	0.75	0.75	0.75	0.01
SMO	100	0	1	1	0	0	0	0	1	1	1	0
IBk	97.95	2.04	0.92	0	0.01	0.1	36.26	77.22	0	0	0	0.02

But when evaluating using the test set, the results were as follows:

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	99.19	0.8	0.99	0.77	0	0.08	18.82	66.46	0.78	0.78	0.78	0
NB	97.74	2.25	0.99	0.61	0.02	0.15	53.32	111.05	0.45	1	0.62	0.02
J48	98.06	1.93	0.73	0.46	0.01	0.13	45.17	102.98	0.47	0.47	0.47	0.01
SMO	98.95	1.04	0.76	0.64	0.01	0.1	24.47	75.79	0.85	0.52	0.64	0.01
IBk	98.78	1.21	0.82	0.51	0.01	0.1	32.05	74.09	1	0.34	0.51	0.01

Conclusion: TL is the winner

1.8 Experimental Stat Results for Setting Number 8:

- Target (Training) Dataset: Size=392 (384 Inactive + 8 Active)
- Testing Dataset: Size=1043 (1024 Inactive + 19 Active)
- In this experiment I did 10 fold cross validation

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	99.48	0.51	0.97	0.87	0	0.07	12.02	50.31	0.87	0.87	0.87	0
NB	97.95	2.04	0.99	0.65	0.01	0.13	44.81	95.51	0.5	1	0.66	0.02
J48	98.97	1.02	0.87	0.74	0.01	0.1	23.95	71.38	0.75	0.75	0.75	0.01
SMO	99.74	0.25	0.93	0.93	0	0.05	5.98	35.69	1	0.87	0.93	0
IBk	99.48	0.51	0.92	0.85	0	0.05	12.09	42.21	1	0.75	0.85	0

But when evaluating using the test set, the results were as follows:

	corr	incorr	auc	kap	mae	rmse	rae	rrse	prec	rec	fM	err rate
TL	99.23	0.76	0.99	0.78	0	0.08	17.28	59.91	0.78	0.78	0.78	0
NB	98.08	1.91	0.99	0.64	0.01	0.13	47.78	103.43	0.48	1	0.65	0.01
J48	99.52	0.47	0.94	0.86	0	0.06	11.91	51.74	0.85	0.89	0.87	0
SMO	99.13	0.86	0.81	0.72	0	0.09	21.45	69.41	0.85	0.63	0.72	0
IBk	98.75	1.24	0.94	0.51	0.01	0.09	30.79	70.77	0.87	0.36	0.51	0.01

Conclusion: TL is the winner