Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Σχολή Μηχανικών
Ελληνικό Μεσογειακό Πανεπιστήμιο

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Πολυ-ομική ανάλυση γενομικών δεδομένων για την κατηγοριοποίηση δειγμάτων γλοιοβλαστώματος

Σπουδαστής:
Νικόλαος Οικονόμου

Αριθμός μητρώου:
ΤΠ4845

Επιβλέπων καθηγητής
Μανώλης Τσικνάκης

Συν-Επιβλέπων καθηγητής
Λευτέρης Κουμάκης

ΗΡΑΚΛΕΙΟ, ΔΕΚΕΜΒΡΙΟΣ 2022

Περιεχόμενα

- Τι μελετάει η εν λόγω πτυχιακή;
- Τι είναι το γλοιοβλάστωμα;
- Μηχανική μάθηση & αλγόριθμοι
- Μεθοδολογίες για την ομαδοποίηση πολλαπλών ομικών δεδομένων
- Μεθοδολογία
- Αποτελέσματα
- Αποτελέσματα σε άλλα καρκινικά δεδομένα
- Συμπεράσματα
- Μελλοντικές επεκτάσεις

Τι μελετάει η εν λόγω πτυχιακή;

Τι μελετάει η εν λόγω πτυχιακή;

- Τα τελευταία χρόνια, παράγονται όλο και περισσότερα βιολογικά δεδομένα λόγω της ραγδαίας εξέλιξης της τεχνολογίας στον χώρο της βιολογίας [1]
- Στο παρελθόν γινόταν η ανάλυση κάθε επιπέδου ξεχωριστά για την ανακάλυψη της πρόκλησης μιας ασθένειας ή εύρεσης μιας θεραπείας
- Οι περισσότερες ασθένειες επηρεάζουν πολύπλοκα μοριακά μονοπάτια όπου διαφορετικά βιολογικά ομικά επίπεδα αλληλεπιδρούν μεταξύ τους [2]
- Με την πτυχιακή αυτή, προσπαθούμε να αναλύσουμε και να κατηγοριοποιήσουμε πολύ-ομικά δεδομένα από ασθενείς με γλοιοβλάστωμα χρησιμοποιώντας την στρατηγική Early integration και μοντέλα μηχανικής μάθησης

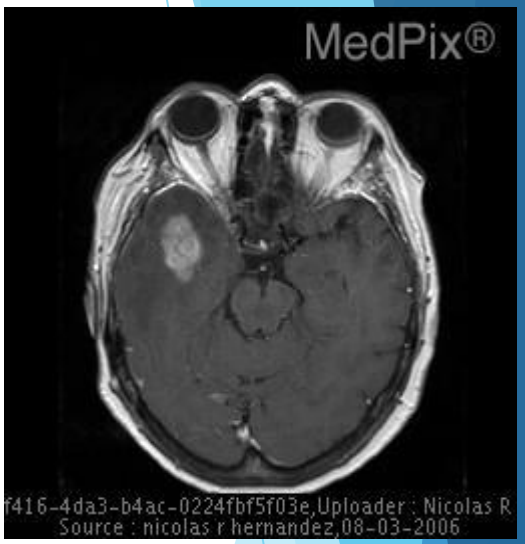
[1] Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. Bioinformatics and biology insights, 14, 1177932219899051. Ανακτήθηκε 8/5/2022 από: https://journals.sagepub.com/doi/full/10.1177/1177932219899051

[2]Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal, 19, 3735-3746. https://doi.org/10.1016/j.csbj.2021.06.030

Τι είναι το γλοιοβλάστωμα;

Τι είναι το γλοιοβλάστωμα;

- Αποτελεί τον πιο κοινό πρωτοπαθή και θανατηφόρο καρκίνο του κεντρικού νευρικού συστήματος [3]
- Η εμφάνιση του είναι πιθανή σε οποιαδήποτε ηλικία αλλά κυρίως παρατηρείται στις ηλικίες 55-60
- Το προσδόκιμο ζωής των ασθενών που διαγιγνώσκονται με γλοιοβλάστωμα να ανέρχεται στους 14-15 μήνες μετά την διάγνωση [4]
- Τα συνήθη συμπτώματα της νόσου είναι πονοκέφαλοι απώλεια μνήμης, σύγχυση, νευρικές διαταραχές ή επιληπτικές κρίσεις. [5]



Πηγή: https://medpix.nlm.nih.gov/case?id=0c6fc001-3971-40b1-857a-6769b015e7f8

[3] Yin, W., Tang, G., Zhou, Q., Cao, Y., Li, H., Fu, X., Wu, Z., & Jiang, X. (2019). Expression Profile Analysis Identifies a Novel Five-Gene Signature to Improve Prognosis Prediction of Glioblastoma. *Frontiers in genetics*, *10*, 419. https://doi.org/10.3389/fgene.2019.00419

[4] Hanif, F., Muzaffar, K., Perveen, K., Malhi, S. M., & Simjee, S. (2017). Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. *Asian Pacific journal of cancer prevention : APJCP*, *18*(1), 3–9. https://doi.org/10.22034/APJCP.2017.18.1.3

[5] Alifieris, C., & Trafalis, D. T. (2015). Glioblastoma multiforme: Pathogenesis and treatment. *Pharmacology & therapeutics*, *152*, 63–82. https://doi.org/10.1016/j.pharmthera.2015.05.005

Μηχανική μάθηση και αλγόριθμοι

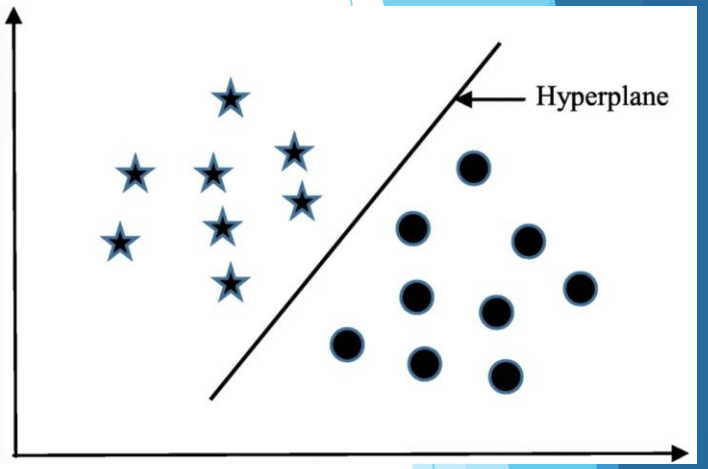
Μηχανική μάθηση & Αλγόριθμοι (1/3)

- Η μηχανική μάθηση είναι ένας κλάδος της επιστήμης υπολογιστών που επιδιώκει να δώσει την δυνατότητα στους υπολογιστές στο να «μάθουν» χωρίς να είναι απευθείας προγραμματισμένοι
- Κάθε φορά που ο υπολογιστής «μαθαίνει», αποκτά εμπειρία καταφέρνοντας έτσι να αυτοβελτιώνεται [6]
- Χωρίζεται συνήθως σε τρεις μεγάλες κατηγορίες
  - Επιτηρούμενη μάθηση (Supervised Learning)
  - Μη Επιτηρούμενη μάθηση (Unsupervised Learning)
  - Ενισχυτική μάθηση (Reinforcement Learning) [7]

[6] Qifang Bi, Katherine E Goodman, Joshua Kaminsky, Justin Lessler, What is Machine Learning? A Primer for the Epidemiologist, American Journal of Epidemiology, Volume 188, Issue 12, December 2019, Pages 2222–2239, https://doi.org/10.1093/aje/kwz189
[7] Mahesh, Batta. (2019). Machine Learning Algorithms -A Review. 10.21275/ART20203995. Ανακτήθηκε 15/4/2022 από https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review

Μηχανική μάθηση & Αλγόριθμοι (2/3)

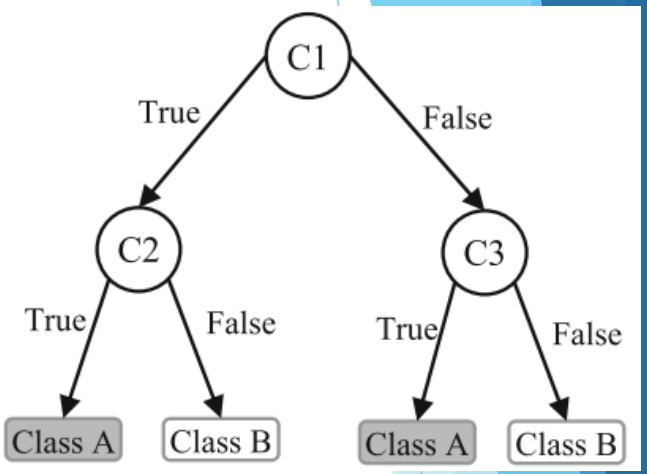
- Ο SVM (Support Vector Machine) είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται σε προβλήματα επιτηρούμενης μάθησης και μπορεί να ταξινομήσει γραμμικά (linear) και μη γραμμικά (non-linear) δεδομένα
- Αναγνωρίζει το υπερπλάνο (hyperplane) το οποίο χωρίζει τα δεδομένα σε δυο κλάσεις
- Μπορεί να διαχειριστεί μεγάλο όγκο δεδομένων
- Είναι αποτελεσματικός σε δεδομένα όπου ο αριθμός των διαστάσεων είναι μεγαλύτερος από τον αριθμό των δειγμάτων



Πηγή: https://link.springer.com/article/10.1186/s12911-019-1004-8

Μηχανική μάθηση & Αλγόριθμοι (3/3)

- Ο Decision Tree αποτελεί έναν από τους παλαιότερους αλγορίθμους μηχανικής μάθησης
- Ταξινομεί τα δεδομένα βάση των τιμών των χαρακτηριστικών οπτικοποιώντας τα σε δενδροειδή μορφή
- Κάθε κόμβος αναπαριστά ένα χαρακτηριστικό (feature) για ταξινόμηση ενώ κάθε κλαδί αναπαριστά μια τιμή που ο κόμβος μπορεί να υποθέσει.
- Η ταξινόμηση αρχίζει από τον κόμβο ρίζα (root node). Τα φύλλα ή τερματικοί κόμβοι αντιστοιχούν στα αποτελέσματα της απόφασης του δέντρου απόφασης.[8]



Πηγή: https://link.springer.com/article/10.1186/s12911-019-1004-8

[8] Uddin, S., Khan, A., Hossain, M. *et al.* Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak* **19**, 281 (2019). https://doi.org/10.1186/s12911-019-1004-8

Μεθοδολογίες για την ομαδοποίηση πολλαπλών ομικών δεδομένων

Μεθοδολογίες για την ομαδοποίηση πολλαπλών ομικών δεδομένων (1/2)

- Συνήθως, για την ανάλυση πολλαπλών ομικών δεδομένων, συγκεντρώνεται και συνδυάζεται όλη η διαθέσιμη πληροφορία από κάθε ομικό επίπεδο (πχ DNA Methylation, miRNA, Gene Expression κ.α)
- Έτσι γίνεται η χρήση όλου του εύρους της πληροφορίας
- Αυτό επιτυγχάνεται με αλγορίθμους μηχανικής μάθησης παράγοντας έτσι βιοδείκτες διάγνωσης και ομαδοποίησης
- Ορισμένες μεθοδολογίες είναι οι:
  - Early integration
  - Mixed integration
  - Intermediate integration

Μεθοδολογίες για την ομαδοποίηση πολλαπλών ομικών δεδομένων (2/2)

- Η στρατηγική Early integration, βασίζεται στην συνένωση κάθε συνόλου δεδομένων (dataset) σε έναν μεγάλο ενιαίο πίνακα
- Αυτό έχει ως αποτέλεσμα την αύξηση των μεταβλητών (columns) με σταθερό όμως αριθμό παρατηρήσεων (rows)
- Χρησιμοποιείται αρκετά συχνά
- «Δυνατό της χαρακτηριστικό» είναι η ικανότητα της να μπορεί να συνδυάσει δεδομένα από κάθε ομικό επίπεδο επιτρέποντας στους αλγορίθμους μηχανικής μάθησης να ανιχνεύσουν συσχετίσεις μεταξύ των διαφορετικών επιπέδων [9]

[9] Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*, *19*, 3735-3746. https://doi.org/10.1016/j.csbj.2021.06.030

Μεθοδολογία

Μεθοδολογία (1/3)

- Αντλήθηκαν δεδομένα ασθενών γλοιοβλαστώματος από το TCGA (DNA Methylation, miRNA, Gene expression, clinical & survival δεδομένα)
- Με τη χρήση βιβλιοθηκών της Python, επιλέχθηκαν οι ασθενείς όπου είχαν το χαρακτηριστικό days_to_last_followup > 100 στο clinical αρχείο
- Κατηγοριοποίηση των ασθενών βάση του χαρακτηριστικού CDE_vital_status, το οποίο φανερώνει αν ο ασθενής είναι (LIVING) ή όχι στη ζωή (DECEASED)
- Προέκυψε μια νέα αναπαράσταση του κλινικού αρχείου που περιλαμβάνει μόνο τους ασθενείς με τα παραπάνω κριτήρια

	sampleID	CDE_vital_status	days_to_last_followup
0	TCGA.02.0001.01	DECEASED	279.0
1	TCGA.02.0003.01	DECEASED	144.0
2	TCGA.02.0004.01	DECEASED	345.0
3	TCGA.02.0006.01	DECEASED	558.0
4	TCGA.02.0007.01	DECEASED	705.0
...	...	...	...
442	TCGA.74.6573.01	DECEASED	105.0
443	TCGA.74.6573.11	DECEASED	105.0
444	TCGA.76.4926.01	DECEASED	138.0
445	TCGA.76.4927.01	DECEASED	535.0
446	TCGA.87.5896.01	LIVING	800.0

447 rows × 3 columns

Κλινικό αρχείο μετά την επεξεργασία

Μεθοδολογία (2/3)

- Βρέθηκαν οι κοινοί ασθενείς μεταξύ των αρχείων
- Λόγω της διακύμανσης των τιμών που παρατηρήθηκε στα δεδομένα, προχωρήσαμε στην διαδικασία της ομαλοποίησης των δεδομένων
- Γίνεται χρήση της μεθοδολογίας Early integration
- Το νέο σύνολο δεδομένων (ενοποιημένο) αποτελείται 230 ασθενείς

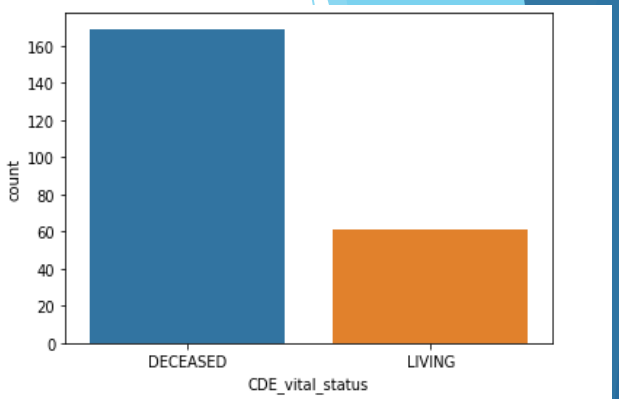
	AACS	FSTL1	ELMO2	CREB3L1	RPS11	PNMA1	MMP2	SAMD4A	SMARCD3	A4GNT	...	cg27622610	cg27626299	cg2762642
TCGA.02.0001.01	0.249831	0.603452	0.145579	0.227189	0.536511	0.281552	0.654650	0.057786	0.160680	1.000000	...	0.634403	0.061604	0.97394
TCGA.02.0003.01	0.259859	0.730249	0.321741	0.216045	0.467085	0.182134	0.745141	0.055900	0.256339	0.375168	...	0.574758	0.262108	0.19256
TCGA.02.0007.01	0.427716	0.152753	0.075147	0.450566	0.333490	0.433460	0.085272	0.106165	0.118726	0.766229	...	0.627240	0.709814	0.93428
TCGA.02.0009.01	0.554233	0.854778	0.423709	0.334683	0.397484	0.171951	0.759044	0.102445	0.198096	0.256758	...	0.404914	0.569276	0.71672
TCGA.02.0010.01	0.637423	0.627497	0.657317	0.514724	0.543211	0.375774	0.094958	0.036727	0.769906	0.688522	...	0.009895	0.039988	0.95304
...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
TCGA.41.2572.01	0.525578	0.829292	0.726965	0.225229	0.532250	0.647362	0.640516	0.383696	0.795864	0.423438	...	0.384363	0.674163	0.88188
TCGA.41.2573.01	0.729782	0.668766	0.825307	0.120989	0.516259	0.814925	0.534055	0.101132	0.861436	0.277688	...	0.180606	0.242916	0.57605
TCGA.41.2575.01	0.501280	0.638301	0.586010	0.228132	0.603503	0.568709	0.589915	0.175332	0.889185	0.397727	...	0.083526	0.196087	0.65677
TCGA.41.3393.01	0.417496	0.726087	0.694750	0.214937	0.518122	0.673899	0.641191	0.414615	0.750362	0.216226	...	0.472694	0.643516	0.89749
TCGA.41.3915.01	0.301592	0.887816	0.476522	0.113209	0.561694	0.655163	0.554971	0.296717	0.650863	0.440657	...	0.354195	0.151799	0.83671

230 rows × 17577 columns

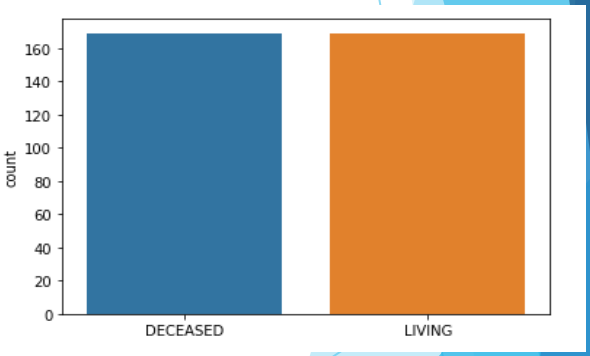
Ενοποιημένο σύνολο δεδομένων

Μεθοδολογία (3/3)

- Παρατηρήθηκε ότι τα νέα σύνολα δεδομένων είναι ανισόρροπα, με την κλάση DECEASED να υπερισχύει της LIVING
- Έγινε εφαρμογή του αλγορίθμου SMOTE μετατρέποντας έτσι τη μειοψηφούσα κλάση όσο την πλειοψηφούσα
- Τα δεδομένα χωρίστηκαν σε Train set (70%) & Test set (30%)
- Έγινε εκπαίδευση των αλγορίθμων SVM & Decision Tree



Πριν την εφαρμογή SMOTE



Μετά την εφαρμογή SMOTE

Αποτελέσματα

Αποτελέσματα (1/6)

- Αρχικά **δεν** χρησιμοποιήθηκε ο αλγόριθμος SMOTE

```
SVM Accuracy: 0.7101449275362319
DecisionTree Accuracy: 0.6231884057971014
```

Gene Expression

```
SVM Accuracy: 0.782608695652174
DecisionTree Accuracy: 0.6086956521739131
```

DNA Methylation

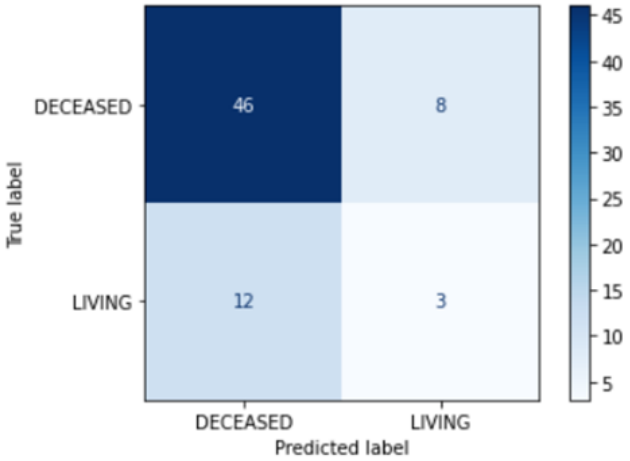
```
SVM Accuracy: 0.5652173913043478
DecisionTree Accuracy: 0.5362318840579711
```

miRNA

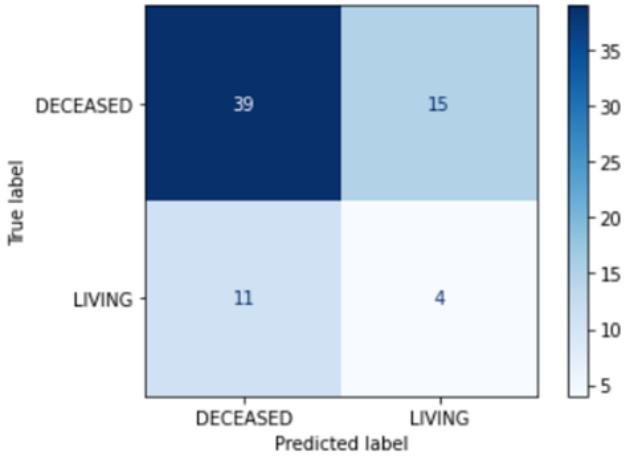
```
SVM Accuracy: 0.6956521739130435
DecisionTree Accuracy: 0.6086956521739131
```

Ενοποιημένο σύνολο

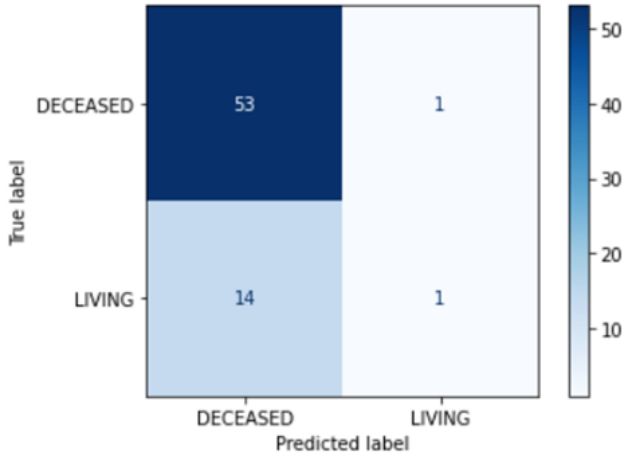
Αποτελέσματα (2/6)



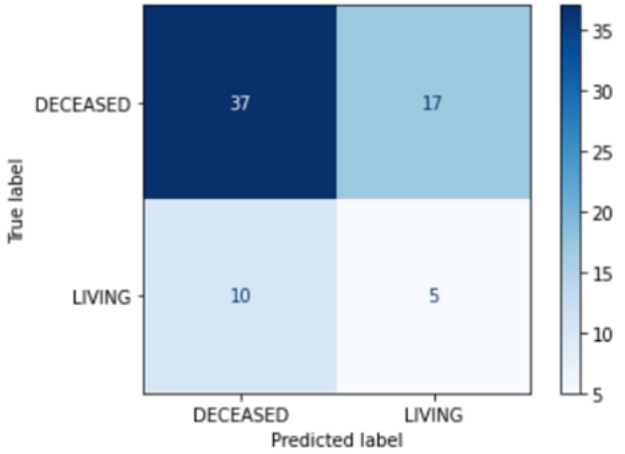
Gene expression - SVM



Gene expression – Decision Tree

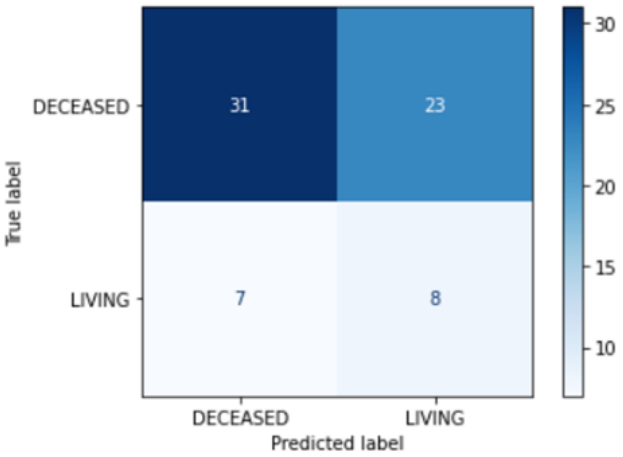


DNA Methylation - SVM

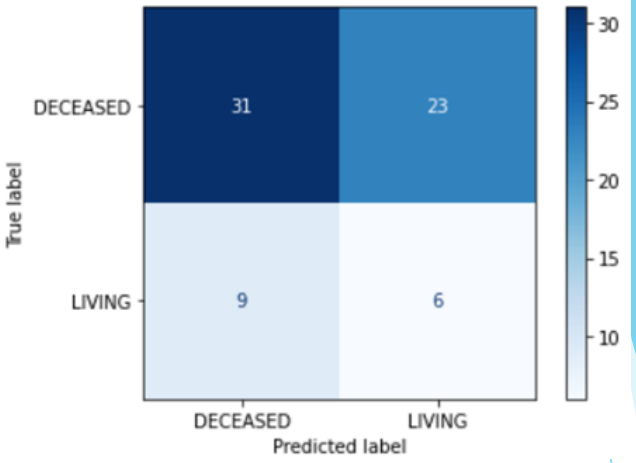


DNA Methylation – Decision Tree

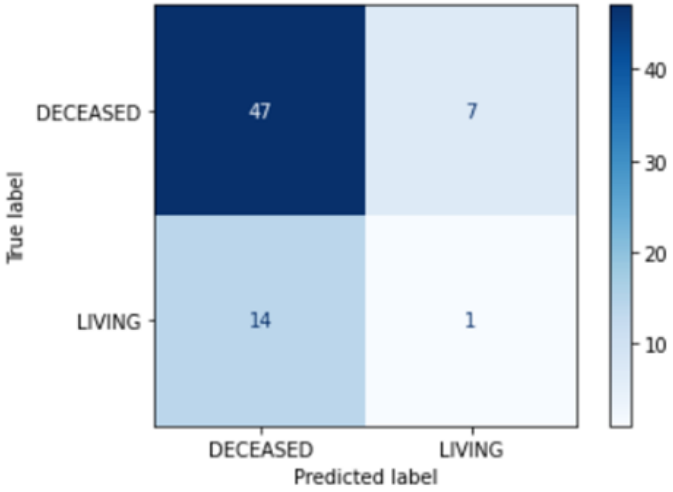
Αποτελέσματα (3/6)



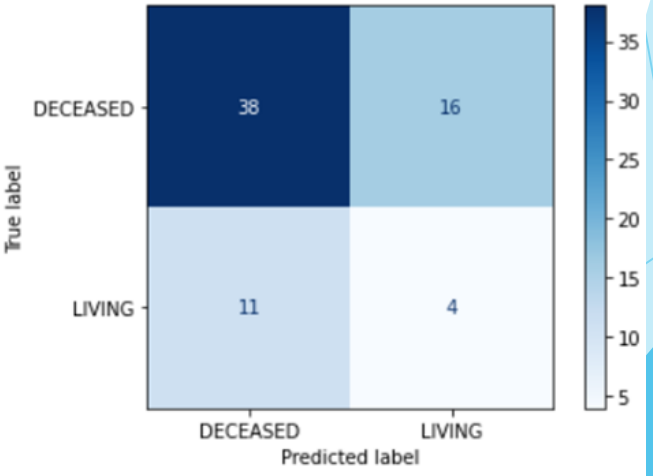
miRNA - SVM



miRNA– Decision Tree



Ενοποιημένο σύνολο- SVM



Ενοποιημένο σύνολο– Decision Tree

Αποτελέσματα (4/6)

- **Γίνεται** χρήση του αλγορίθμου SMOTE στα δεδομένα

```
SVM Accuracy: 0.9019607843137255
DecisionTree Accuracy: 0.7156862745098039
```

Gene Expression

```
SVM Accuracy: 0.8823529411764706
DecisionTree Accuracy: 0.6862745098039216
```

DNA Methylation

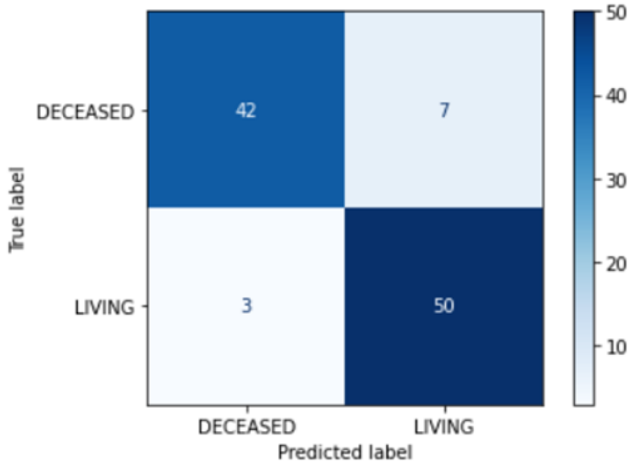
```
SVM Accuracy: 0.7352941176470589
DecisionTree Accuracy: 0.7352941176470589
```

miRNA

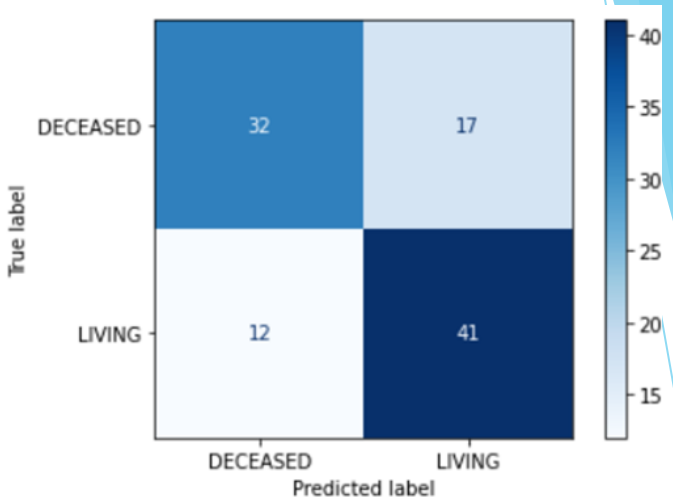
```
SVM Accuracy: 0.8823529411764706
DecisionTree Accuracy: 0.7254901960784313
```

Ενοποιημένο σύνολο

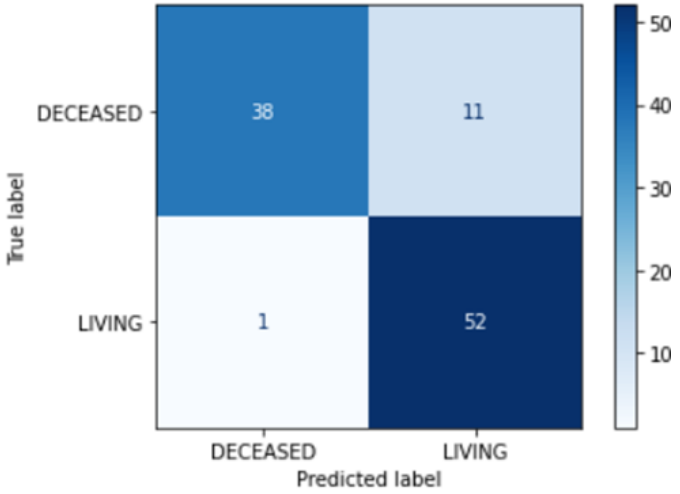
Αποτελέσματα (5/6)



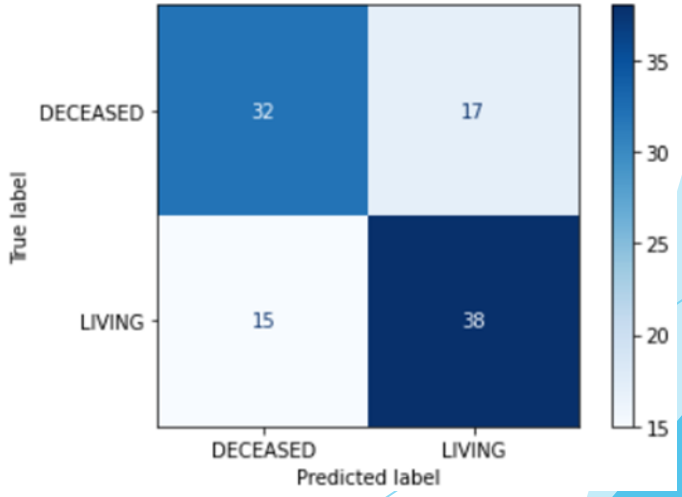
Gene expression - SVM



Gene expression – Decision Tree

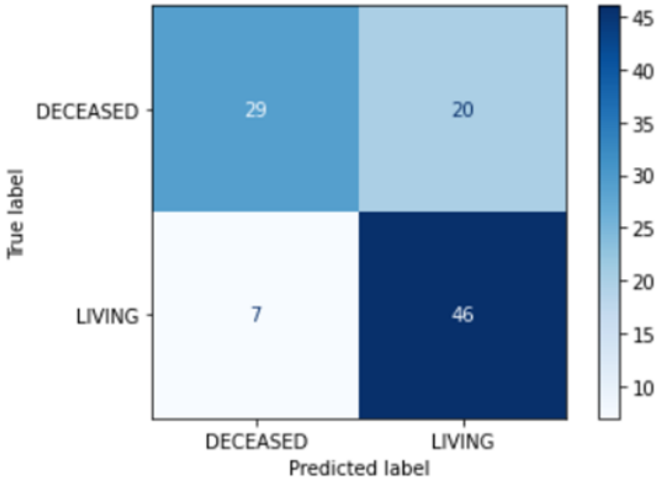


DNA Methylation - SVM

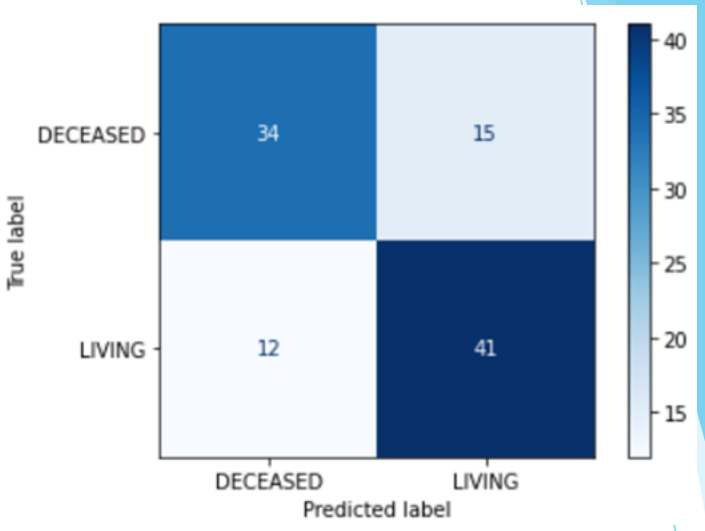


DNA Methylation – Decision Tree

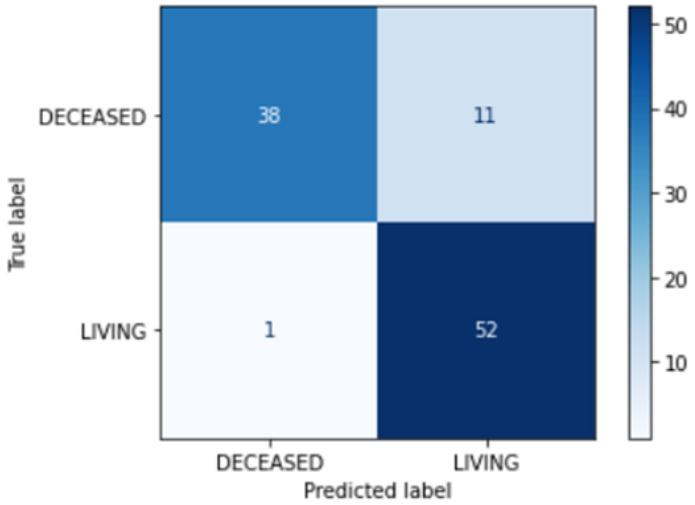
Αποτελέσματα (6/6)



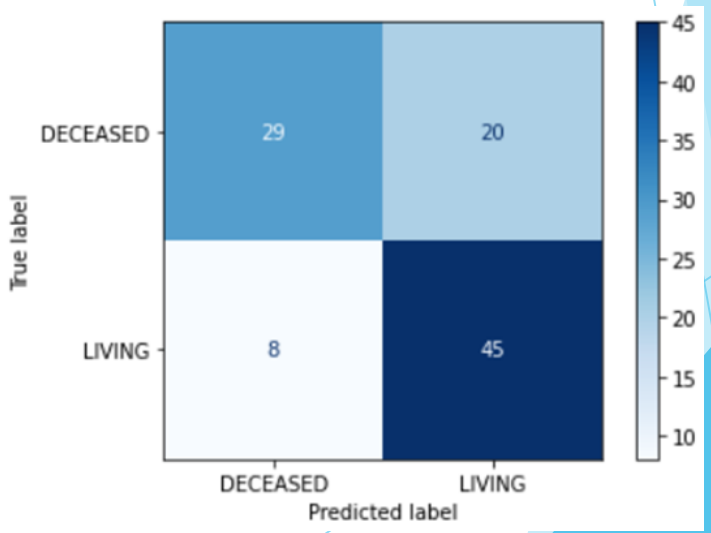
miRNA - SVM



miRNA– Decision Tree



Ενοποιημένο σύνολο- SVM



Ενοποιημένο σύνολο– Decision Tree

Αποτελέσματα σε άλλα καρκινικά δεδομένα

Αποτελέσματα σε άλλα καρκινικά δεδομένα (1/3)

- Αντλήθηκαν από το TCGA δεδομένα ασθενών από 9 διαφορετικούς τύπους καρκίνου.
- Ακολουθήθηκαν οι ίδιες σχεδόν μεθοδολογίες που χρησιμοποιήθηκαν και στην επεξεργασία των δειγμάτων γλοιοβλαστώματος
- Έγινε χρήση του αλγορίθμου επιλογής χαρακτηριστικών ANOVA
- Έγινε εκπαίδευση των μοντέλων SVM και Decision Tree με τα νέα χαρακτηριστικά

Αποτελέσματα σε άλλα καρκινικά δεδομένα (2/3)

Αποτελέσματα ακρίβειας πρόβλεψης **πριν** την χρήση του αλγορίθμου ANOVA

Cancer	SVM (Gene Expression)	Decision Tree (Gene Expression)	SVM (DNA Methylation)	Decision Tree (DNA Methylation)	SVM (miRNA)	Decision Tree (miRNA)	SVM (Ενοποιημένο σύνολο)	Decision Tree (Ενοποιημένο σύνολο)
LIHC	63.93%	60.65%	59.83%	54.91%	59.83%	61.47%	63.93%	54.91%
BIC	88.70%	82.25%	88.17%	76.34%	90.32%	79.56%	89.24%	80.64%
OV	49.42%	48.27%	58.62%	56.32%	56.32%	56.32%	55.17%	51.72%
SARC	59.49%	64.55%	63.29%	56.96%	55.69%	58.22%	64.55%	58.22%
KIRC	69.35%	61.29%	72.58%	70.96%	66.12%	64.51%	74.19%	54.83%
LUSC	47.05%	53.92%	52.94%	55.88%	51.96%	50%	49.01%	45.09%
COAD	75.38%	69.23%	72.30%	67.69%	73.84%	64.61%	72.30%	70.69%
AML	64.58%	39.58%	60.41%	50%	56.25%	58.33%	64.58%	54.16%
SKCM	58.33%	50.75%	54.54%	53.78%	53.03%	56.06%	56.06%	51.51%

Αποτελέσματα σε άλλα καρκινικά δεδομένα (3/3)

Αποτελέσματα ακρίβειας πρόβλεψης **μετά** την χρήση του αλγορίθμου ANOVA

Cancer	SVM (Gene Expression)	Decision Tree (Gene Expression)	SVM (DNA Methylation)	Decision Tree (DNA Methylation)	SVM (miRNA)	Decision Tree (miRNA)	SVM (Ενοποιημένο σύνολο)	Decision Tree (Ενοποιημένο σύνολο)
LIHC	**72.95%**	**67.21%**	**62.29%**	52.45%	**71.31%**	54.91%	**70.49%**	52.45%
BIC	**91.93%**	**83.87%**	**91.93%**	**78.49%**	**91.93%**	**80.10%**	**91.93%**	**84.40%**
OV	**59.77%**	**55.17%**	**68.96%**	**62.06%**	**66.66%**	**64.36%**	**74.71%**	**66.66%**
SARC	**63.29%**	**65.82%**	**65.82%**	53.16%	**67.08%**	**62.02%**	62.02%	**62.02%**
KIRC	**79.03%**	**72.58%**	**82.25%**	58.06%	**85.48%**	**66.12%**	**82.25%**	**75.80%**
LUSC	**63.72%**	**61.76%**	**60.78%**	55.88%	**60.78%**	**50.98%**	**61.76%**	**52.94%**
COAD	**76.92%**	66.15%	**75.38%**	**72.30%**	**78.46%**	63.07%	**76.92%**	56.92%
AML	**75%**	**68.75%**	**81.25%**	**75%**	**72.91%**	**62.5%**	**77.08%**	**60.41%**
SKCM	**67.42%**	46.96%	**61.36%**	**55.30 %**	**61.36%**	51.51%	**66.66%**	**61.36%**

Συμπεράσματα

Συμπεράσματα (1/2)

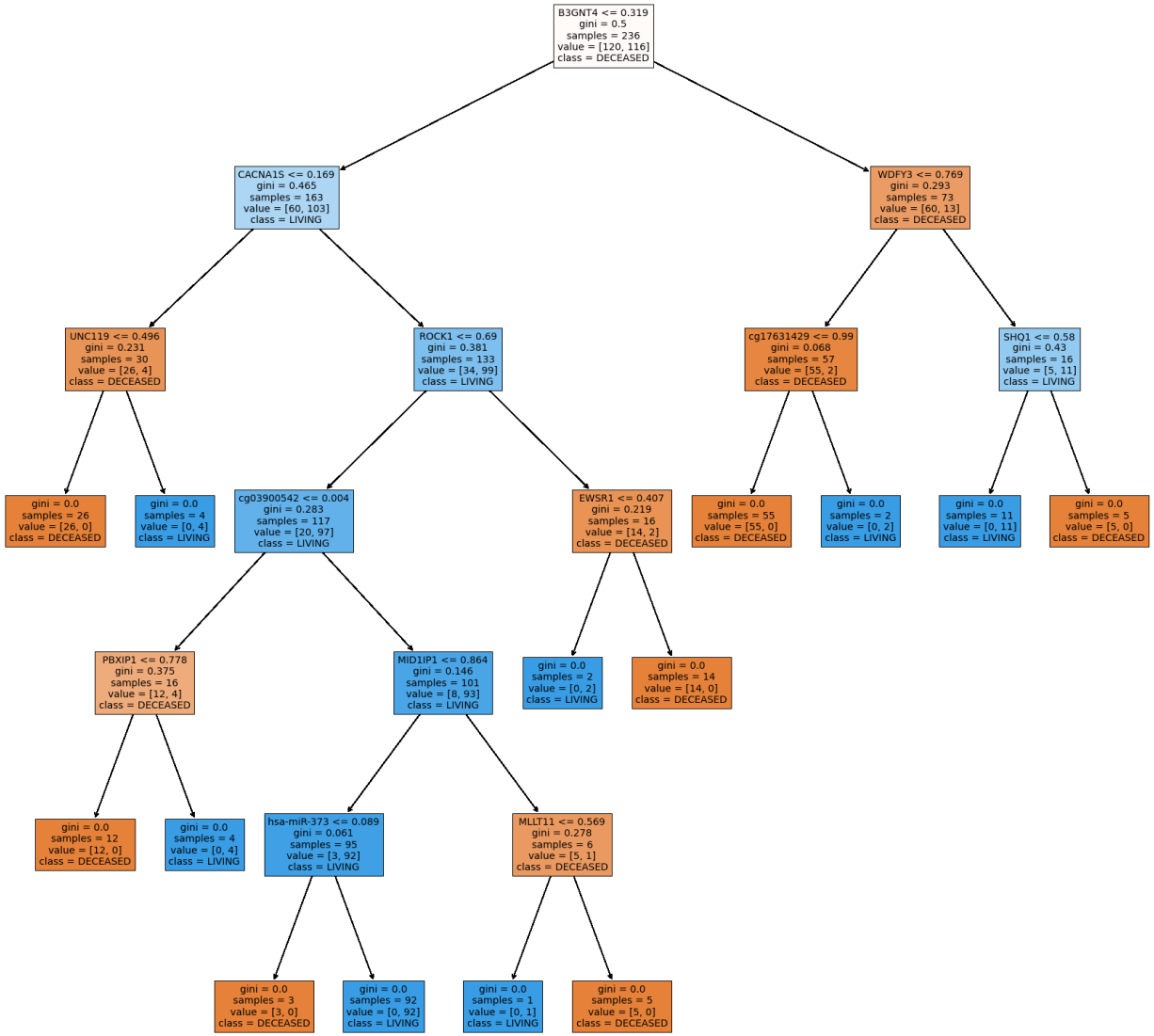
- Ο αλγόριθμος SVM είχε σχεδόν πάντα καλύτερη κατηγοριοποίηση των δεδομένων γλοιοβλαστώματος που διαχειριστήκαμε
- Με και χωρίς την χρήση του αλγορίθμου SMOTE
- Από το δέντρο απόφασης με τη χρήση SMOTE βρέθηκαν χαρακτηριστικά απ' όλα τα ομικά επίπεδα (Gene expression, miRNA, DNA Methylation)
- Η στρατηγική early integration, έχει πράγματι καλύτερα αποτελέσματα σε αντίθεση με την μεμονωμένη ανάλυση
- Κάποια από τα χαρακτηριστικά που σχετίζονται με την ασθένεια & πιθανή θεραπεία του γλοιοβλαστώματος είναι:

  - MLLT11
  - PBXIP1
  - ROCK1
  - miR-373

Γονίδια

miRNA

Συμπεράσματα (2/2)



Δέντρο απόφασης από το ενοποιημένο σύνολο δεδομένων

Μελλοντικές επεκτάσεις

Μελλοντικές επεκτάσεις

- Μελλοντική επέκταση της πτυχιακής εργασίας θα μπορούσε να ήταν:
  - Επιλογή διαφορετικού χαρακτηριστικού για την κατηγοριοποίηση των ασθενών από το κλινικό αρχείο
  - Χρήση μοντέλου νευρωνικού δικτύου αντί τους αλγορίθμους μηχανικής μάθησης
  - Έτσι ίσως παρατηρηθεί αν παρουσιάζεται μεγαλύτερη ακρίβεια στην πρόβλεψη και νέα αποτελέσματα στα δεδομένα του γλοιοβλαστώματος που χρησιμοποιήσαμε

Ευχαριστώ για την προσοχή σας!
Ερωτήσεις;