

# MELANOMA MACHINE LEARNING

Σαλτερης Γεωργιος 2136  
Μουλόπουλος Αντώνιος 2104

Αποτελέσματα απο την εκπαίδευση διαφορων classifiers που προβλεπουν αν καποιος εχει μελανομα η όχι. Οι classifier που χρησιμοποιουμε είναι οι Random Forest, Linear Regression,SVM,Linear Logistic Regression,OLS και GaussianNB.

Για το παρακατω πειραμα κρατήσαμε ολα τα snr με **pvalue**  $\leq p = 0.0001$ .

Οι ασθενεις μας ειναι στο συνολο 4980. Απο αυτους οι 3962 δεν εχουν την ασθeneia(**control** = 0) και οι υπόλοιποι 1018 εχουν την ασθeneia(**case** = 1).

Τα snrs με **pvalue**  $\leq p = 0.0001$  είναι 5415 για να αποφύγουμε το **overfitting** μειώνουμε τα snrs με τον εξής τρόπο, τον οποίο σκεφτήκαμε μόνοι μας. Αφού στήσουμε τον πίνακα με του ασθενείς και τα χαρακτηριστικά τους, στη συνέχεια βρίσκουμε τον πίνακα με το **cosine similarity** μεταξύ των snrs χρησιμοποιώντας την εντολή **metrics.pairwise.cosine\_similarity**. Ορίζουμε έναν counter για κάθε snr. Στη συνέχεια εξετάζουμε την ομοιότητα μεταξύ των snrs και αυξάνουμε κατά ένα τον counter του snr που έχει τιμή στον πίνακα ίση η μικρότερη από 0.3(οσο πιο κοντά στο 1 είναι η τιμή τόσο πιο όμοια είναι μεταξύ τους τα χαρακτηριστικά.) κατα 1. Τέλος κρατήσαμε τα snrs των οποίων ο counter είναι μεγαλύτερος ή ίσος από το 60% των συνολικών snrs.

Καναμε ενα τυχαίο sample του παραπανω δειγματος όπου το 90% ειναι το trainData και το 10% ειναι το testData. Μετα το sample στο 90% του δειγματος οι 3572 ασθενεις δεν εχουνε την ασθeneia( και 910 ασθενεις εχουν την ασθeneia. Στο 10% του δειγματος οι 390 ασθενεις δεν εχουν την ασθeneia και οι 108 εχουν την ασθeneia.

	Cross validation	AUC	recall	precision
Random Forest	0,84	0,64	0,28	0,22
Linear Regression	0,89	0,81	0,63	0,6
SVM	0,78	0,5	0,0	0,0
SVM(kernel = linear)	0,92	0,85	0,71	0,74
Linear Logistic Regression	0,92	0,87	0,75	0,68
OLS	0,91	0,82	0,73	0,62
Naive Bayes Gaussian	0,96	0,93	0,91	0,89

Στη συνέχεια θα κανουμε ένα balanced του Data Train. Το αρχικό Data Train έχει 3572 ασθενεις που δεν εχουνε την ασθeneia και 910 ασθενεις που εχουν την ασθeneia. Θα μειώσουμε τους ασθενεις που δεν εχουν την ασθeneia στους 1300 με τυχαίο τροπο ώστε να υπάρχει μια ισορροπια μεταξυ των ασθενών που εχου την ασθeneία και αυτών που δεν την έχουν.

	<b>Cross validation</b>	<b>AUC</b>	<b>recall</b>	<b>precision</b>
<b>Random Forest</b>	<b>0,7</b>	<b>0,63</b>	<b>0,34</b>	<b>0,25</b>
<b>Linear Regression</b>	<b>0,81</b>	<b>0,77</b>	<b>0,61</b>	<b>0,58</b>
<b>SVM</b>	<b>0,65</b>	<b>0,6</b>	<b>0,3</b>	<b>0,13</b>
<b>SVM(kernel = linear)</b>	<b>0,88</b>	<b>0,88</b>	<b>0,89</b>	<b>0,75</b>
<b>Linear Logistic Regression</b>	<b>0,87</b>	<b>0,87</b>	<b>0,88</b>	<b>0,76</b>
<b>OLS</b>	<b>0,81</b>	<b>0,82</b>	<b>0,85</b>	<b>0,6</b>
<b>Naive Bayes Gaussian</b>	<b>0,94</b>	<b>0,94</b>	<b>0,90</b>	<b>0,92</b>