Applicazione di algoritmi di Machine Learning ad una serie di oli essenziali e le loro composizioni chimiche, per sviluppare modelli di classificazione di relazione composizione-attività di tipo quantitativo (QCAR). Analisi e validazione predittiva.

SAPIENZA
UNIVERSITÀ DI ROMA

Facoltà di Farmacia e Medicina

Corso di Laurea in Farmacia

Tesi Sperimentale in Chimica Farmaceutica

A.A. 2021/2022

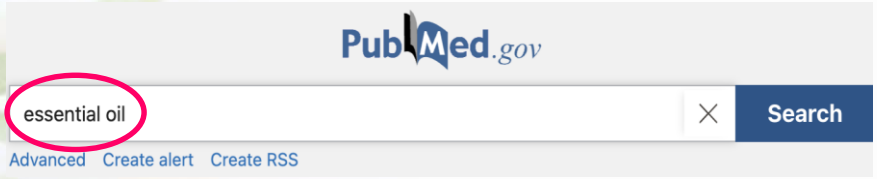
Laureanda: Roberta Astolfi

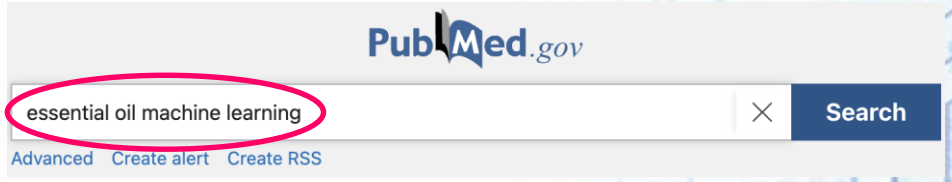
Matricola: 1766291

Relatore: Prof. Rino Ragno

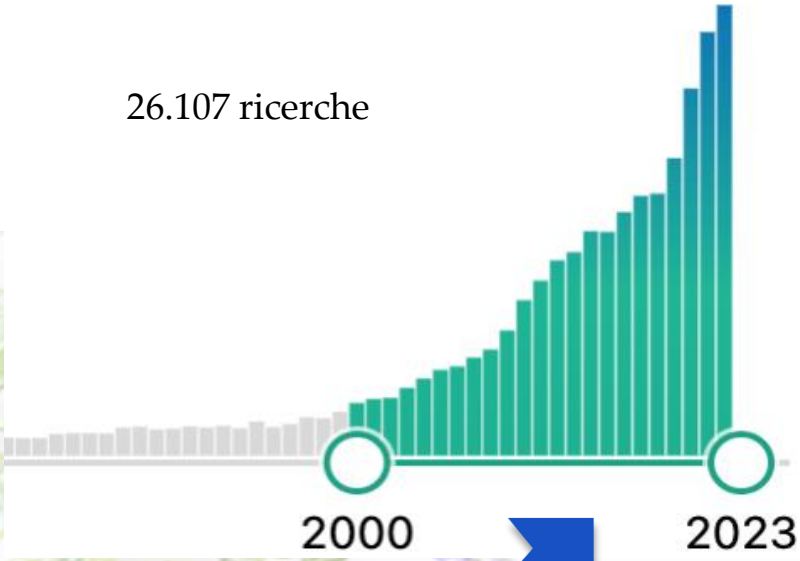
Correlatrice: Dott.ssa Alessandra Oliva

Scopo della tesi

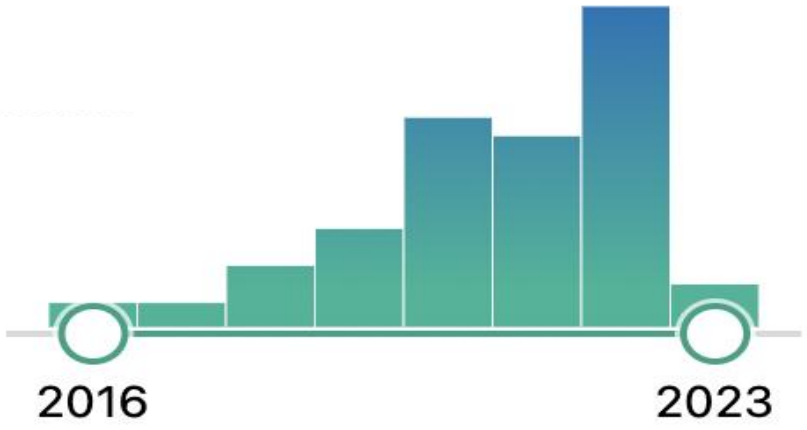




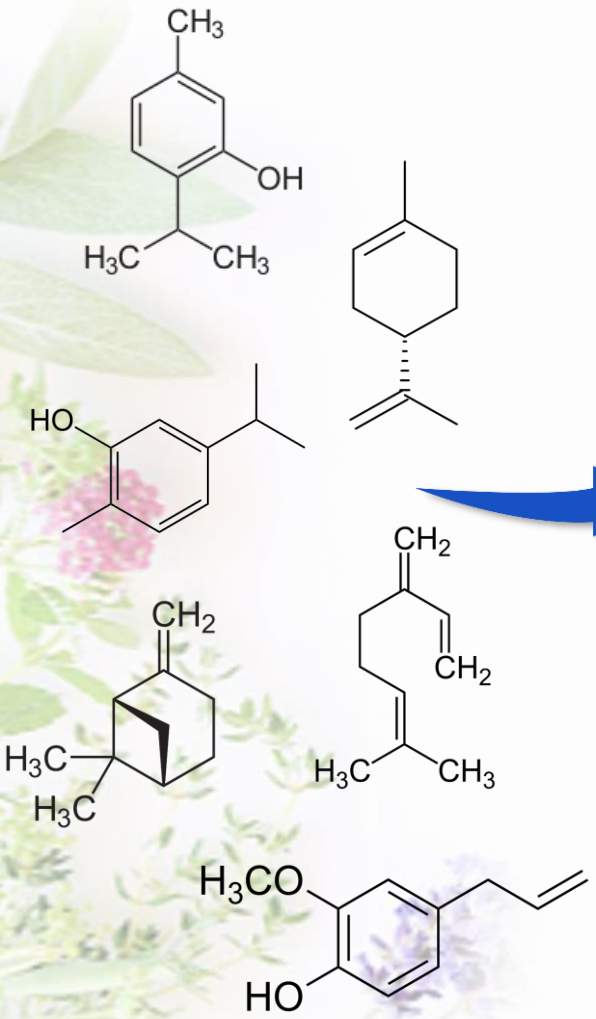
26.107 ricerche



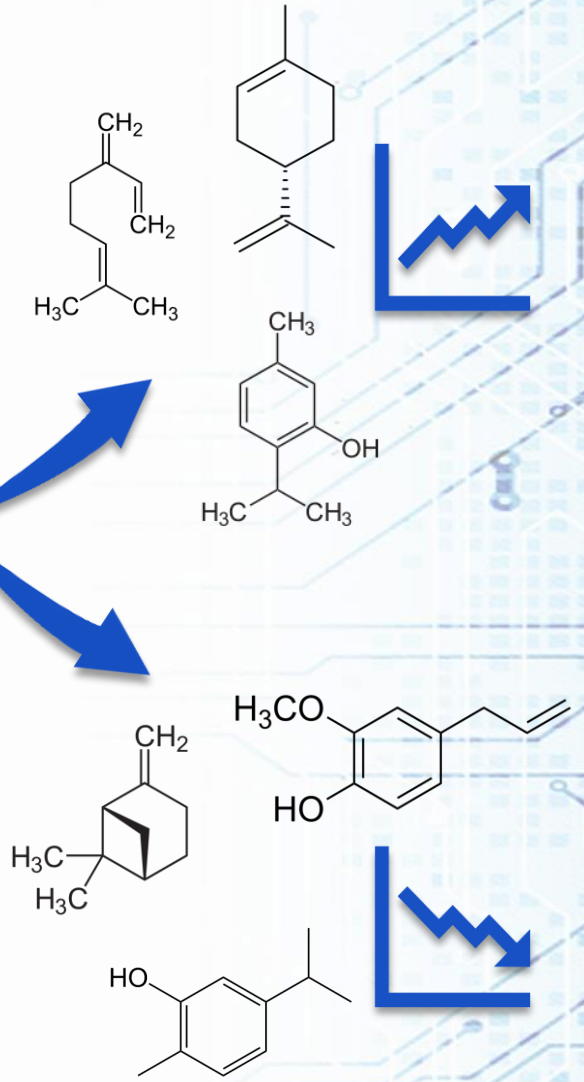
40 ricerche



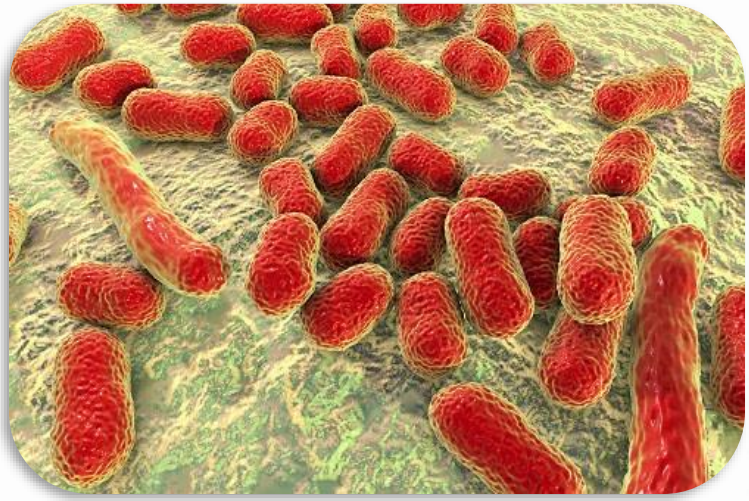
Scopo della tesi

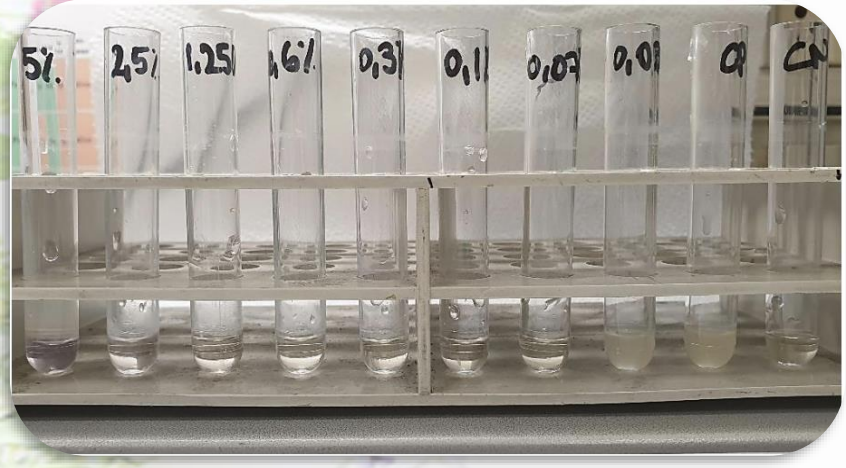


Machine Learning

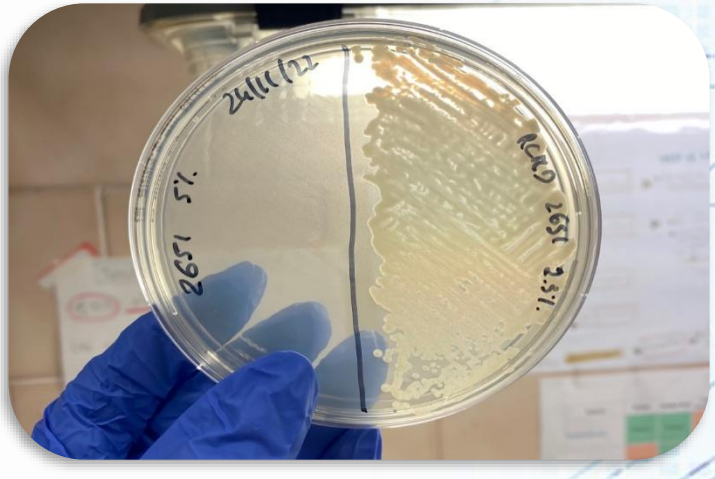


Acinetobacter baumannii





MIC



MBC

Protocollo Sperimentale

Stats

2648	1081	1183	20547
Extracts compositions	Plants	Publications analysed	Biological activities stored

Activity ID	Assay Type	Target Organism	Target Name	Target Code	Incubation Time	Activity Type	Fixed Doses Value	Activity Mode	Activity Value
2220	Functional	Bacteria G+	Staphylococcus aureus			MIC		Undefined	1.25 %
2221	Functional	Bacteria G+	Methicillin Resistant Staphylococcus Aureus (MRSA)			MIC		Undefined	0.63 %
2222	Functional	Bacteria G-	Pseudomonas Aeruginosa			MIC		Undefined	5.0 %
2223	Functional	Bacteria G+	Staphylococcus aureus			MBC		Undefined	5.0 %
2224	Functional	Bacteria G+	Methicillin Resistant Staphylococcus Aureus (MRSA)			MBC		Undefined	1.25 %
2225	Functional	Bacteria G-	Pseudomonas Aeruginosa			MBC		Undefined	5.0 %
2226	Functional	Bacteria G-	Acinetobacter Baumannii			MIC		Undefined	1.25 %
2227	Functional	Bacteria G-	Acinetobacter Baumannii			MBC		Undefined	1.25 %
2228	Functional	Bacteria G-	Klebsiella pneumoniae Carbapenemase			MIC		Undefined	2.5 %
2229	Functional	Bacteria G-	Klebsiella pneumoniae Carbapenemase			MBC		Undefined	5.0 %

Protocollo Sperimentale

61 OE
247 componenti

103 OE
460 componenti

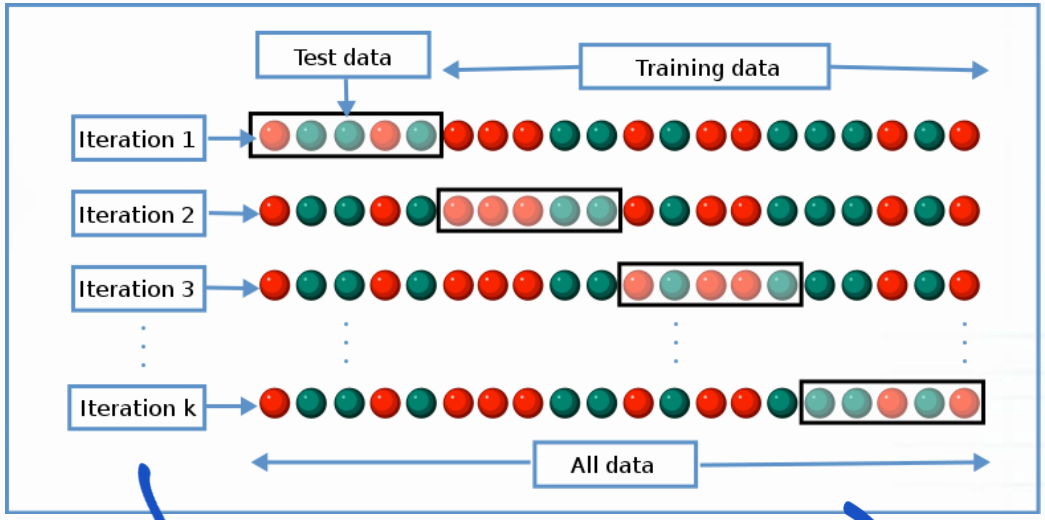
Data Cleaning

Random Search

- SVM
- Gradient Boosting
- Logistic Regression
- Random Forest
- Decision Tree
- K-Nearest Neighbour

Sostanza attiva	SMILES	InChI
Acido acetico	CC(=O)O	1S/C2H4O2/c1-2(3)4/h1H3,(H,3,4)

Random Search




```
################################## Threshold MIC = 0.03 ==> Actives 21 and Inactives 82 ##############################
###
Threshold: 0.03; Classifier: gb ; random_iters: 10000; nlevel: 0; scaling: False; pc: 0.7; comp_shape: (103, 349); X_
shape: (103, 11); mcc_fit:  0.683; mcc_cv:  0.534
Threshold: 0.03; Classifier: gb ; random_iters: 10000; nlevel: 0; scaling: False; pc: 0.8; comp_shape: (103, 349); X_
shape: (103, 16); mcc_fit:  1.000; mcc_cv:  0.485
Threshold: 0.03; Classifier: gb ; random_iters: 10000; nlevel: 0; scaling: False; pc: 0.9; comp_shape: (103, 349); X_
shape: (103, 25); mcc_fit:  0.879; mcc_cv:  0.246
Threshold: 0.03; Classifier: gb ; random_iters: 10000; nlevel: 0; scaling: False; pc: 0.99; comp_shape: (103, 349); X
_shape: (103, 44); mcc_fit:  0.879; mcc_cv:  0.272
Threshold: 0.03; Classifier: gb ; random_iters: 10000; nlevel: 1; scaling: False; pc: 0.7; comp_shape: (103, 349); X_
shape: (103, 11); mcc_fit:  0.683; mcc_cv:  0.491
Threshold: 0.03; Classifier: gb ; random_iters: 10000; nlevel: 1; scaling: False; pc: 0.8; comp_shape: (103, 349); X_
shape: (103, 16); mcc_fit:  1.000; mcc_cv:  0.573
```

Modelli Ottimizzati – I Training Set

THRESHOLD MIC = 0.03% v/v	
CLASSIFICATORE	SVM
ITERAZIONI RANDOM SEARCH	1 000 000
N°LIVELLI	3
SCALING	FALSE
PCA	0.9
MCC_CV	0.631
MCC_FIT	0.729

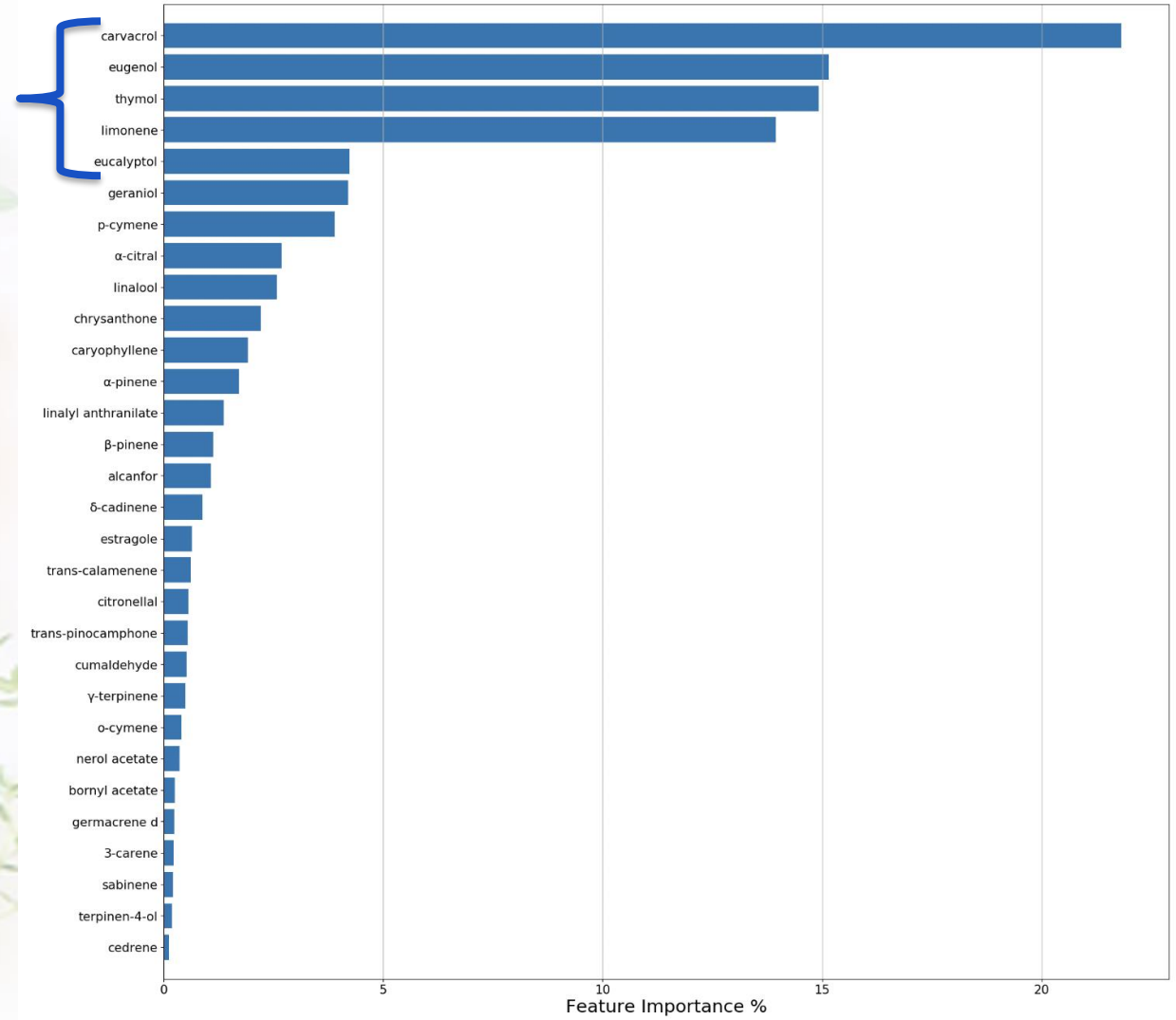
THRESHOLD MBC = 0.03% v/v	
CLASSIFICATORE	SVM
ITERAZIONI RANDOM SEARCH	1 000 000
N°LIVELLI	1
SCALING	FALSE
PCA	0.9
MCC_CV	0.728
MCC_FIT	0.821

Modelli Ottimizzati – II Training Set

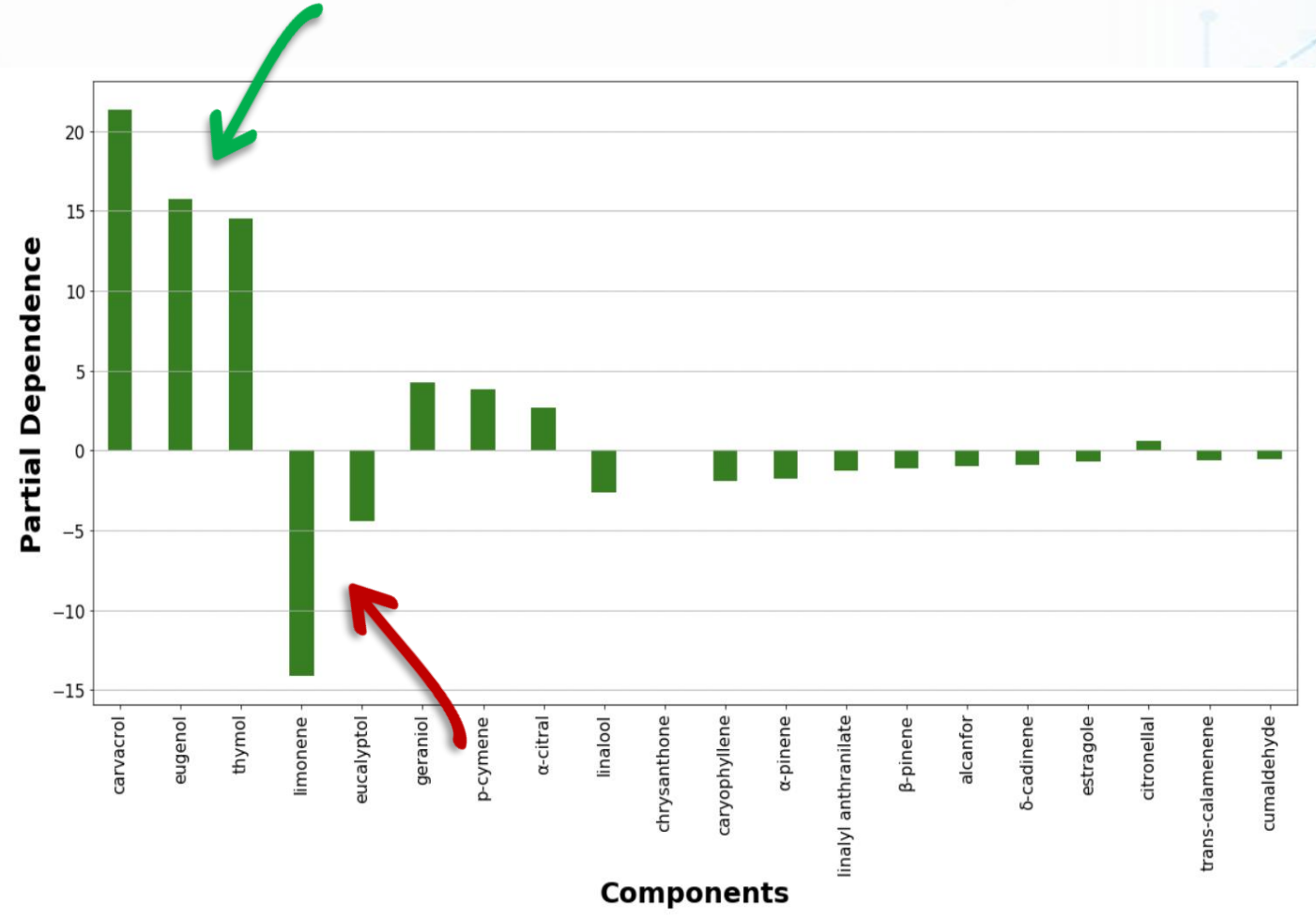
THRESHOLD MIC = 0.03% v/v	
CLASSIFICATORE	GB
ITERAZIONI RANDOM SEARCH	1 000
N°LIVELLI	0
SCALING	FALSE
PCA	0.7
MCC_CV	0.677
MCC_FIT	0.739

THRESHOLD MBC = 0.03% v/v	
CLASSIFICATORE	LR
ITERAZIONI RANDOM SEARCH	1 000 000
N° LIVELLI	3
SCALING	FALSE
PCA	0.7
MCC_CV	0.723
MCC_FIT	0.723

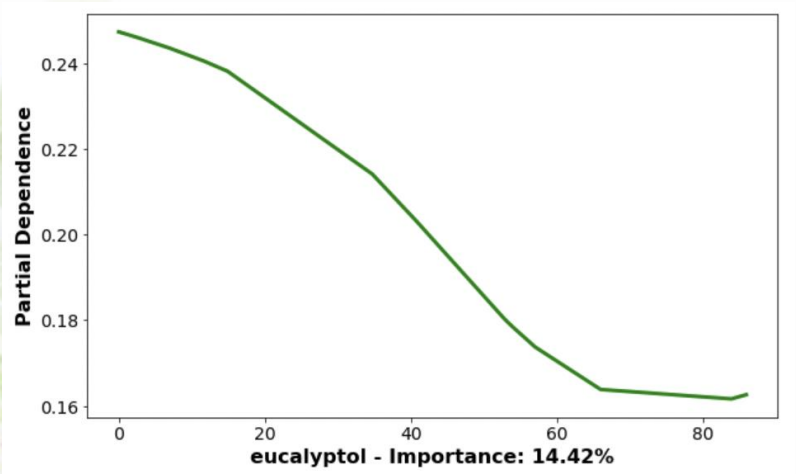
Analisi Skater MIC – I Training Set

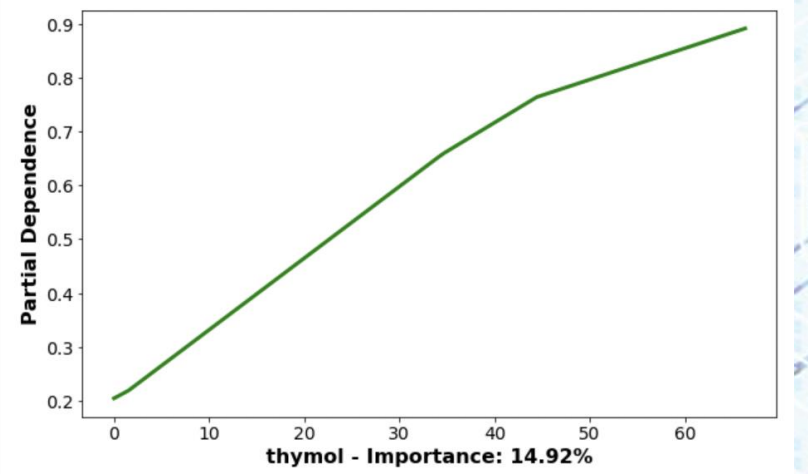


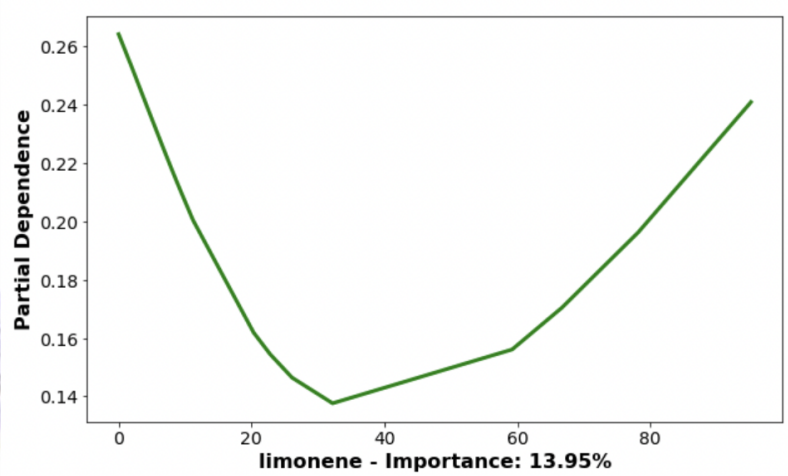
Partial Dependence MIC – I Training Set



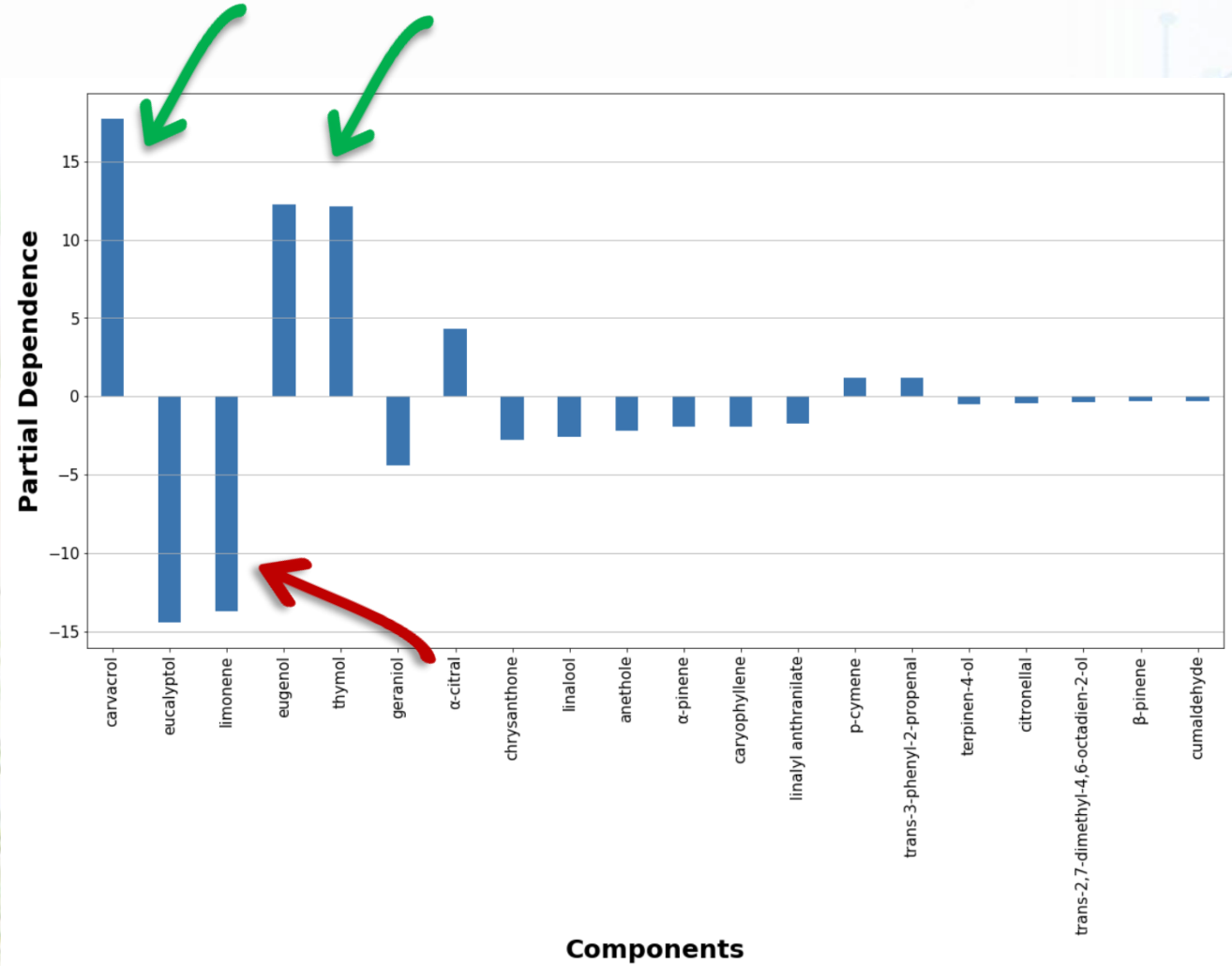
Partial Dependence MIC – I Training Set



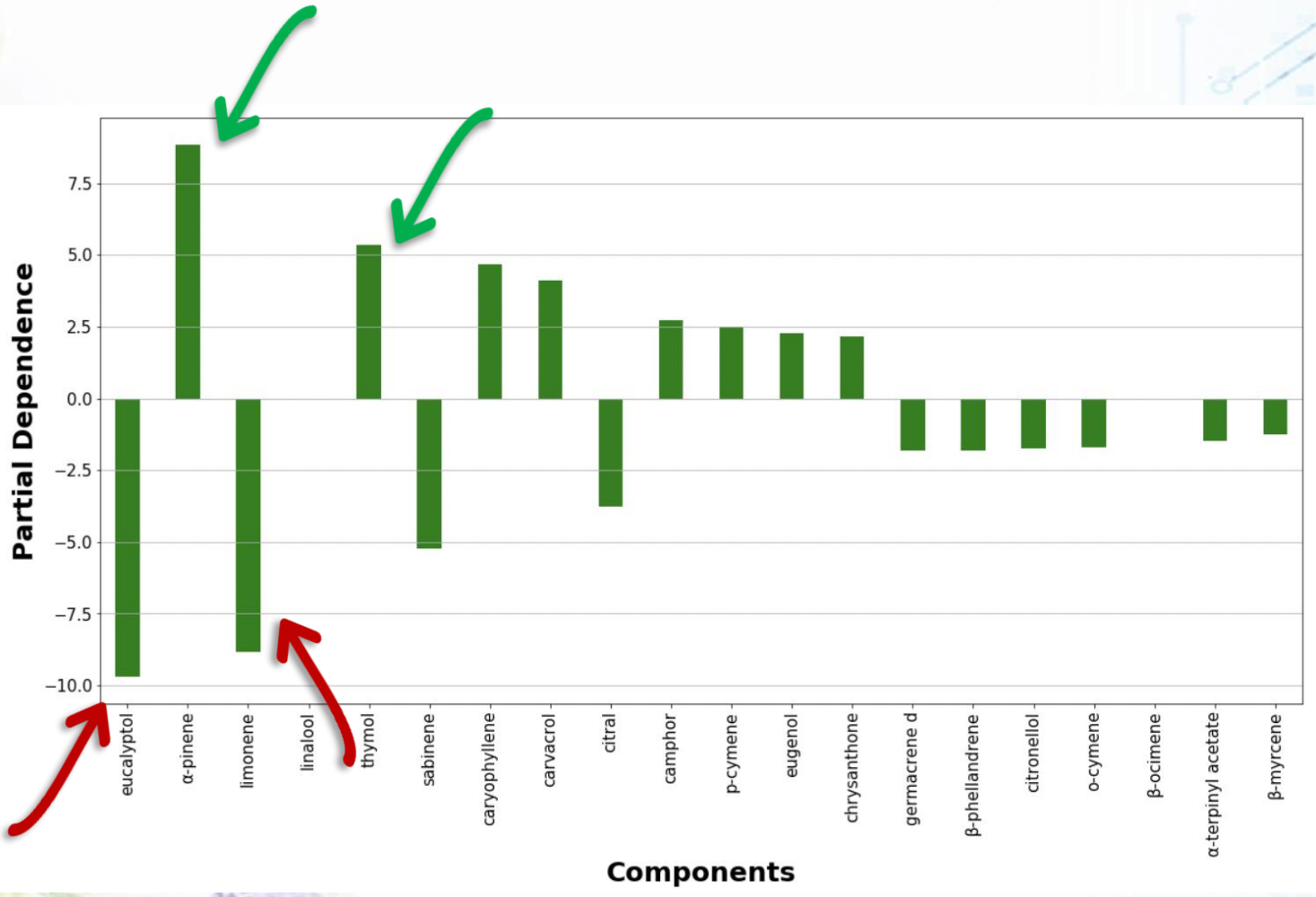




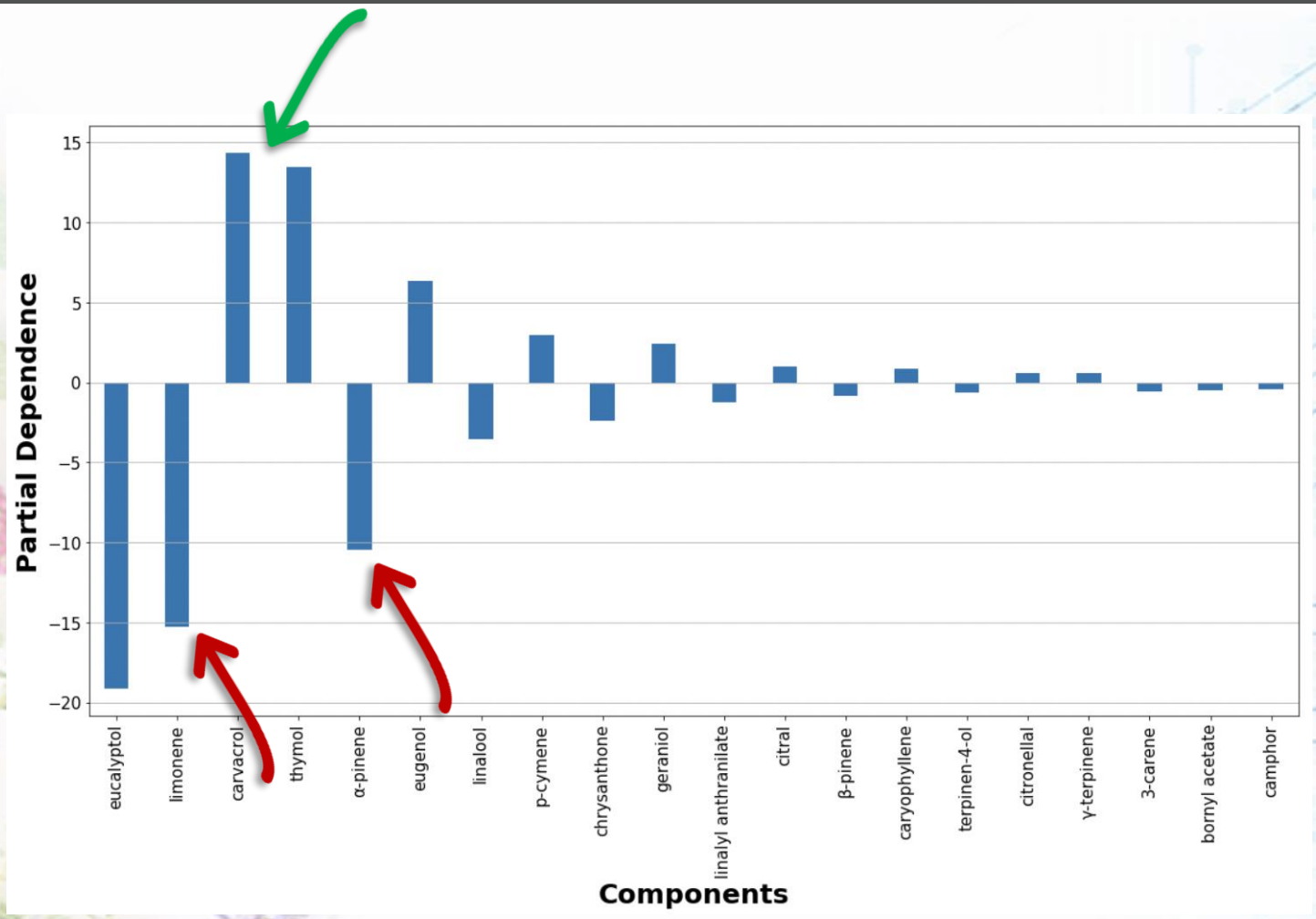
Partial Dependence MBC – I Training Set



Partial Dependence MIC – II Training Set



Partial Dependence MBC – II Training Set



Test Set – Predizione Attività

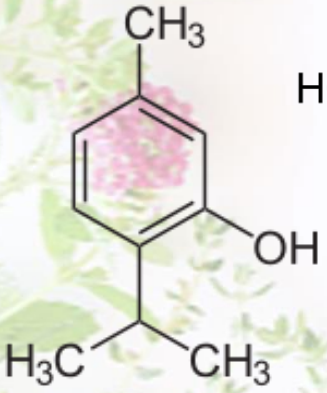
NUMERO OLIO ESSENZIALE	PIANTA DI PROVENIENZA
RCMD 2090	*Foeniculum vulgare* - Finocchio
RCMD 2634	*Anethum graveolens* - Aneto
RCMD 2642	*Thymus vulgaris* - Timo
RCMD 2643	*Mentha L.* - Menta
RCMD 2647	*Rosmarinus officinalis* - Rosmarino
RCMD 2651	*Origanum majorana* - Maggiorana
RCMD 2678	*Thymus vulgaris* – Timo ct. Thujanolo
RCMD 2679	*Laurus nobilis* - Alloro
RCMD 2680	*Cupressus L.* - Cipresso
RCMD 2683	*Helyichrysum* - Elicriso
RCMD 2685	*Lavandula angustifolia* - Lavanda

Analisi Microbiologica

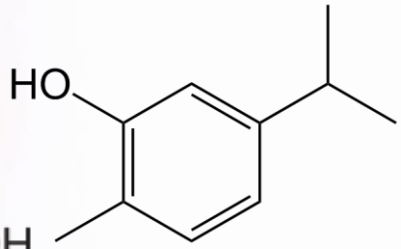
NUMERO OLIO ESSENZIALE	PIANTA DI PROVENIENZA	MIC (% v/v)	MBC (% v/v)
RCMD 2090	*Foeniculum vulgare* - Finocchio	2,5%	2,5%
RCMD 2634	*Anethum graveolens* – Aneto	> 5%	n.e
RCMD 2642	*Thymus vulgaris* – Timo	> 5%	n.e
RCMD 2643	*Mentha L.* – Menta	> 5%	n.e
RCMD 2647	*Rosmarinus officinalis* - Rosmarino	> 5%	n.e
RCMD 2651	*Origanum majorana* - Maggiorana	5%	5%
RCMD 2678	*Thymus vulgaris* – Timo ct. Thujanolo	1,25%	1,25%
RCMD 2679	*Laurus nobilis* – Alloro	> 5%	n.e
RCMD 2680	*Cupressus L.* – Cipresso	> 5%	n.e
RCMD 2683	*Helyichrysum* – Elicriso	> 5%	n.e
RCMD 2685	*Lavandula angustifolia* - Lavanda	5%	5%

Conclusioni

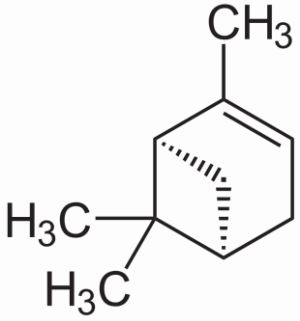
$$\frac{10}{11} = 0.91 \times 100 = 91\%$$



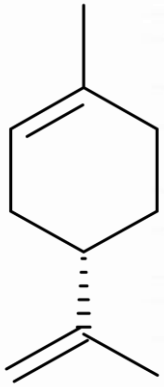
Timolo



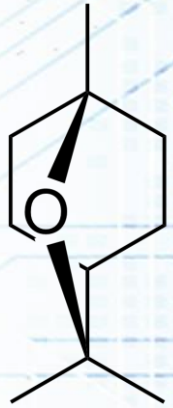
Carvacrolo



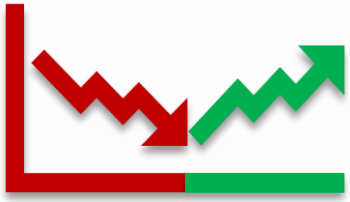
α-Pinene



Limonene



Eucaliptolo



Ringraziamenti

- Prof. Rino Ragno, relatore di tesi
- Dott.ssa Alessandra Oliva, correlatrice di tesi
- Le colleghe di Microbiologia: Dania ed Elisa
- Il dottorando Filippo Sapienza
- La mia compagna di tesi, Giulia

GRAZIE A TUTTI PER L'ATTENZIONE