UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE E TECNOLOGIE

DIPARTIMENTO DI INFORMATICA
GIOVANNI DEGLI ANTONI

Corso di Laurea Magistrale in Informatica

ANALISI DI DATI MULTI-OMICI PER LA PREDIZIONE DELLA PROGNOSI DI PAZIENTI ONCOLOGICI

Relatore: Prof. Elena Casiraghi
Correlatore: Prof. Dario Malchiodi

Tesi di Laurea di:
Alessandro Beranti
Matr. Nr. 977702

ANNO ACCADEMICO 2021-2022

to do

Ringraziamenti

to do

Indice

Ringraziamenti **ii**

Indice **iii**

Introduzione **1**

1 Stato dell'arte **2**
1.1 Apprendimento automatico 2
1.1.1 Apprendimento supervisionato 3
1.1.1.1 Decision Tree 4
1.1.1.2 Random Forest 5
1.1.2 Apprendimento non supervisionato 6
1.1.3 Apprendimento semi-supervisionato 6
1.1.4 Apprendimento per rinforzo 6
1.2 Apprendimento automatico in bioinformatica 7
1.2.1 Apprendimento automatico per la previsione di varianti non codificanti associate a malattie 9
1.2.2 Apprendimento automatico per la previsione del rischio di COVID-19 11

2 Dataset **13**
2.1 Dati Multi-Omici 13
2.1.1 The Cancer Genome Atlas (TCGA) 14
2.1.1.1 Proteine 15
2.1.1.2 mRNA 16
2.1.1.3 miRNA 16
2.1.1.4 CNV 16
2.1.1.5 Etichetta 17

3 Esperimenti **18**
3.1 Preprocessing 18

3.1.1 Scalare i dati . . . 18
3.2 Feature selection . . . 18
3.2.1 Tecniche univariate . . . 19
3.2.1.1 Bassa variabilità . . . 19
3.2.1.2 Mann-Whitney . . . 19
3.2.2 Tecniche multivariate . . . 19
3.2.2.1 Minimum Redundancy Maximum Relevance: mrmr 19
3.2.2.2 Boruta . . . 19
3.2.3 Dimensionalità intrinseca . . . 19
3.2.3.1 ID_twoNN . . . 19
3.2.4 Maximal information-based nonparametric exploration (MINE) 19
3.2.4.1 The maximal information coefficient (MIC) . . . 19
3.2.5 Spearman . . . 19
3.3 Feature extraction . . . 19
3.3.1 Uniform Manifold Approximation: umap . . . 20
3.3.2 t-SNE . . . 20
3.4 Model selection . . . 20
3.4.1 Tuning degli iperparametri . . . 20
3.5 Cross Validation . . . 20
3.6 Metrica di performance . . . 20
3.6.0.0.1 Accuratezza . . . 20
3.6.0.0.2 Marice di confusione . . . 21
3.6.1 Dati sbilanciati . . . 21
3.6.2 Area sotto la curva precision-recall . . . 21
3.7 Risultati . . . 21
3.8 Tecnologie usate . . . 21

4 Conclusioni e sviluppi futuri **22**

Bibliografia **29**

Introduzione

Durante il periodo di tesi mi sono concentrato su

Capitolo 1

Stato dell'arte

1.1 Apprendimento automatico

Il termine apprendimento automatico, comunemente chiamato *Machine learning* in inglese, sta a indicare la capacità dei computer di apprendere e adattarsi agli input forniti. Molto spesso tale termine viene utilizzato per intendere l'intelligenza artificiale e viceversa ma non sono la stessa cosa: il *Machine learning* è un sottoinsieme della categoria più ampia chiamata appunto Intelligenza artificiale.

L'intelligenza artificiale è il campo di computer, sistemi e robot che sono in grado di simulare il comportamento umano in modi che imitano e spesso vanno oltre le capacità umane. I programmi di intelligenza artificiale sono in grado di analizzare e fornire dati o attivare automaticamente azioni senza il bisogno dell'uomo. Alcuni esempi sono gli: assistenti virtuali, anche chiamati *chatbot*, ovvero agenti software in grado di eseguire azioni o erogare servizi in base a comandi ricevuti in maniera vocale o testuale. Questi sistemi, utilizzati sempre di più nel *Customer Care* aziendale come primo livello di assistenza con il cliente, si contraddistinguono per la loro capacità di comprensione del tono del dialogo e di memorizzazione delle informazioni raccolte; sistemi di raccomandazione che indirizzano le scelte degli utenti in base a informazioni da forniti da essi, famosi sono i sistemi che suggeriscono un acquisto in base a quelli fatti precedentemente; il *Natural Language Processing* (NLP) è quel ramo dell'intelligenza artificiale che riguarda l'informazione espressa nel linguaggio naturale. Si tratta di soluzioni che elaborano il linguaggio, con finalità che possono variare dalla comprensione del contenuto, alla traduzione, fino alla produzione di testo in modo autonomo a partire da dati o documenti forniti in input; infine citiamo la *Computer Vision*, area dell'intelligenza artificiale che si occupa dell'analisi e della comprensione di immagini e video con l'obiettivo di estrarre informazioni utili. Ciò può includere il riconoscimento di oggetti, persone,

scene, lettura di testo o la misurazione di proprietà geometriche. Utilizza tecniche di apprendimento automatico, analisi di immagini e algoritmi per elaborare e comprendere i dati visivi. Viene utilizzata in molte applicazioni come la guida autonoma, la videosorveglianza, la diagnostica medica e la realtà aumentata.

Il *Machine learning* utilizza algoritmi per apprendere in maniera automatica intuizioni e riconoscere modelli a partire da dati forniti in input. Gli algoritmi di apprendimento automatico vengono divisi in quattro categorie distinte a seconda del tipo di dato usato per eseguire la fase di apprendimento, queste categorie sono le seguenti:

- apprendimento supervisionato,
- apprendimento non supervisionato,
- apprendimento non supervisionato,
- apprendimento per rinforzo.

1.1.1 Apprendimento supervisionato

Nell'apprendimento supervisionato si ha un insieme x di N osservazioni $x_1, x_2, ..., x_N$ di un vettore avente p dimensioni che contengono esempi di come x sia in relazione con la variabile di output y, anche chiamata etichetta. Usando modelli matematici e statistici adattati ai dati di addestramento, x in questo caso, si vuole cercare di predire l'output y, per dati "nuovi", ovvero dati che il modello non ha usato nella fase di addestramento. L'approccio usato dall'apprendimento supervisionato per "imparare" è quello di estrapolare la relazione che sussiste tra x e y, ovvero imparare usando osservazioni reali. Esistono diversi modi per valutare quanto il modello è riuscito a "imparare bene" e ci danno una stima di quando sia stato in grado di generalizzare, l'argomento verrà approfondito in 3.6.

Il tipo di output che si ricerca influenza il tipo di problema che si sta affrontando, se abbiamo un output numerico siamo di fronte a un problema di regressione mentre se abbiamo un output categorico siamo di fronte a un problema di classificazione. Una variabile numerica possiede un ordine naturale, se prendiamo un'istanza della variabile siamo in grado di dire se sia più grande o piccola di un'altra istanza della stessa variabile. Una variabile numerica può essere rappresentata da un numero reale continuo come da un numero discreto. Le variabile categoriche invece sono sempre discrete e sono prive di un ordine.

Esistono diversi algoritmi di apprendimento supervisionato, ognuno dei quali possiede delle caratteristiche peculiari, in questo lavoro mi concentrerò sui *Random forest*, poiché sono quelli che ho usato nel mio lavoro di tesi e di conseguenza sugli alberi di decisione poiché i primi non sono altro che una foresta dei secondi.

1.1.1.1 Decision Tree

Gli alberi decisionali sono uno dei metodi più comuni usati per eseguire apprendimento supervisionato. Possono essere utilizzati sia per risolvere problemi di regressione che di classificazione, in particolare in quest'ultima trova maggiore applicazione pratica. Un albero di decisione possiede una struttura ad albero ed è composta da:

- nodi non terminali: rappresentano un test su uno o più attributi,
- ramo: rappresenta un esito del test,
- foglia: rappresenta una possibile classe

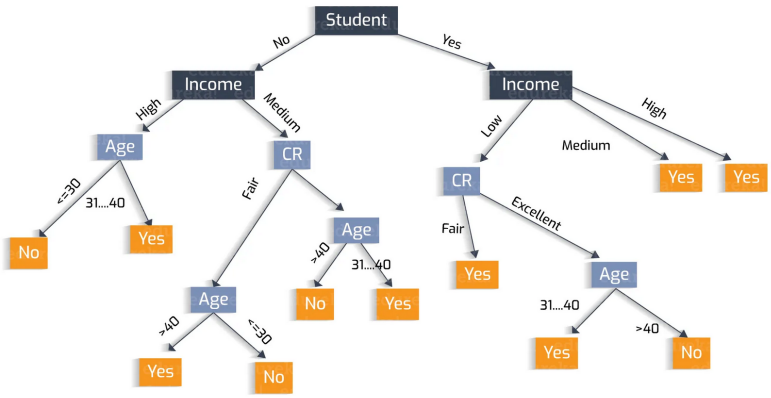


Figura 1: Esempio di albero di decisione

La costruzione di un albero di decisione parte individuando la variabile di predizione più importante che viene utilizzata per la suddivisione del primo nodo interno. Questo processo viene ripetuto per ogni sotto-albero fino al raggiungimento di una condizione di stop:

- tutti i campioni di un nodo appartengono alla stessa classe,
- non ci sono più attributi per un ulteriore partizionamento,
- non rimangono più tuple nelle foglie.

La partizione risultante a seguito di una divisione usando l'attributo selezionato dovrebbe essere pura, ovvero quando tutte le tuple di una data partizione appartengono alla stessa classe. Per scegliere l'attributo su cui effettuare uno *split* vengono comunemente utilizzati due criteri: la Gini impurity e l'entropia.

La *Gini impurity* misura la probabilità che un elemento scelto a caso appartenente a una data classe sia etichettato in modo errato. Un valore basso indica una maggiore purezza delle classi risultanti dallo split. L'entropia è una misura dell'incertezza associata alle classi. Anche in questo caso, un valore basso indica una maggiore purezza delle classi risultanti dallo split. Il criterio di scelta può essere utilizzato per valutare tutti gli attributi disponibili e scegliere quello che produce lo *split* con il valore più basso. In generale si usano questi criteri per evitare *overfitting* e aiutare a generalizzare meglio il modello su dati sconosciuti. Tutte queste tecniche sono euristiche e riescono a trovare un albero efficiente che però non corrisponde all'ottimo, la ricerca dell'albero ottimale è computazionalmente difficile perché il numero di possibili alberi cresce molto rapidamente con il numero di attributi. Prendiamo, per esempio, un insieme di dati con un numero di attributi pari a n che ha solo due valori possibili per ogni attributo, ci sono così 2^n possibili alberi di decisione. Se invece ci fossero stati m valori possibili per ogni attributo, allora ci sarebbero stati m^n possibili alberi di decisione.

La ricerca dell'albero ottimo necessita quindi l'esplorazione di tutte queste possibilità e successivamente la valutazione di quale sia l'albero che ha il minor errore sui dati di addestramento. Per questo motivo vengono usati algoritmi *greedy* che creano l'albero prendendo decisioni localmente ottime usando la *Giny Impurity* o l'entropia.

1.1.1.2 Random Forest

Random forest è una tecnica di apprendimento automatico che si basa sull'utilizzo di molti alberi di decisione per creare un modello più robusto con una precisione migliore rispetto a usare un singolo albero. Esistono diversi motivi per cui si preferisce usare random foresti piuttosto che un singolo albero di decisione:

- miglioramento della precisione: usando più alberi di decisione di solito la precisione finale del modello aumenta,
- riduzione dell'*overfitting*: i singoli alberi di decisione possono soffrire di *overfitting*, fenomeno in cui il modello si adatta troppo bene ai dati usati nell'addestramento e non riesce a generalizzare bene sui dati non usati nell'addestramento. Random Forest utilizza una serie di alberi di decisione e prende la decisione finale come voto di maggioranza, riducendo così l'effetto dell'*overfitting*,
- rilevamento di feature importanti: permette di individuare quali sono le feature più importanti per classificare i dati,
- scalabiltà: scala molto bene con dataset grandi.

1.1.2 Apprendimento non supervisionato

Nell'apprendimento non supervisionato si ha un insieme di N osservazioni $x_1, x_2, ..., x_N$ di un vettore avente p dimensioni ma, al contrario dell'apprendimento supervisionato, l'insieme di dati non ha un'etichetta, i dati sono quindi non annotati. Nel contesto dell'apprendimento supervisionato ci sono diverse metriche che mi indicano quanto il modello è stato in grado di "imparare". Nel contesto dell'apprendimento non supervisionato non esiste una vera e propria misura diretta del successo di un algoritmo. È molto difficile accertare la validità delle inferenze tratte dall'output e dipendono dal tipo di problema e dall'algoritmo utilizzato. A questo scopo si deve ricorrere a euristiche per valutare la qualità dei risultati ottenuti. È importante notare che in genere vengono utilizzate più metriche per valutare l'apprendimento non supervisionato in modo da avere una valutazione completa.

1.1.3 Apprendimento semi-supervisionato

L'apprendimento semi-supervisionato è un tipo di apprendimento automatico in cui dei dati che vengono passati all'algoritmo solamente una piccola parte è etichettata. L'obiettivo dell'algoritmo è quello di usare i dati etichettati per riuscire a fornire una etichetta automaticamente anche ai dati che ne sono privi. Esistono diverse tecniche per fare ciò, come ad esempio le tecniche di propagazione delle etichette [1], le tecniche di co-training [2], le tecniche di self-training [3].

1.1.4 Apprendimento per rinforzo

L'apprendimento per rinforzo è un tipo di apprendimento automatico in cui troviamo un agente che interagisce con uno specifico ambiente in modo da ottenere una ricompensa. L'obiettivo è quello di imparare a prendere decisioni e azioni massimizzando la ricompensa. L'apprendimento per rinforzo si compone di diverse fasi:

- l'agente inizialmente agisce in maniera casuale non avendo nessuna conoscenza dell'ambiente,
- interagendo con l'ambiente l'agente inizia a prendere decisioni e ottenere così le ricompense,
- ottenendo le ricompense l'agente "capisce" come deve agire e adatta la sua *policy*, l'obiettivo è sempre massimizzare la ricompensa.

L'agente utilizza una funzione di valore per valutare le azioni e aiutarlo a scegliere quale azione prendere in un determinato contesto. Esistono diverse funzioni che vengono comunemente usate: funzione di valore Q, funzione di valore V e funzione di valore di *policy* [4].

1.2 Apprendimento automatico in bioinformatica

L'apprendimento automatico viene utilizzato in tantissimi ambiti di ricerca [5]: analisi predittiva e processo decisionale intelligente, sicurezza informatica, *Internet of things* e città smart, previsione del traffico, trasporto pubblico, *healthcare*, pandemia dovuta al COVID-19, NLP, ecc. Questa tecnologia trova spazio anche all'interno della bioinformatica, ovvero l'applicazione di tecniche di calcolo per acquisire e interpretare i dati biologici. È un campo interdisciplinare tra informatica, matematica, statistica, biologia e genetica. La crescita esponenziale di dati biologici disponibili ha sollevato due problemi: un modo efficiente di immagazzinare e gestire dati di questo tipo e riuscire a sfruttarli riuscendo a estrarre informazioni utili. I domini in cui vengono utilizzate tecniche di machine learning sono molte: in Figura 2 vengono mostrate le principali aree nel quale si utilizzano metodi computazionali. All'interno troviamo sei diversi domini: genomica, proteomica, microarray, biologia dei sistemi, evoluzione e text mining. Inoltre troviamo una categoria denominata "other application" nel quale troviamo i restanti problemi.

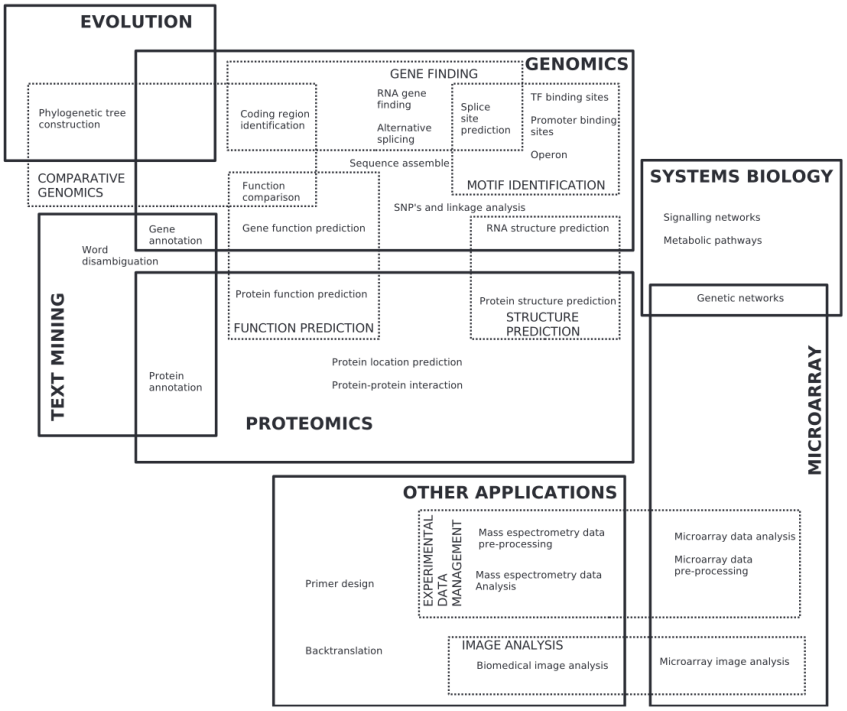


Figura 2: Classificazione dei campi in cui vengono applicati metodi di apprendimento automatico in bioinformatica

La genomica è una delle aree più importanti in bioinformatica. Il numero di sequenze nucleotidiche e proteiche disponibili è in continuo aumento come visibile in figura 3. GenBank è una banca dati a libero accesso e senza restrizione, istituita nel 1982, che riporta tutte le sequenze di nucleotidi e le relative proteine ottenute dopo la loro traduzione. GenBank riceve le proprie informazioni dai risultati ottenuti su oltre 300.000 distinti organismi da laboratori sparsi in tutto il mondo, rappresentando il più importante punto di riferimento nel suo campo di ricerca. A febbraio 2020, conteneva oltre 216 milioni di loci [1] e oltre 399 miliardi di basi da più di 216 milioni di sequenze riportate.

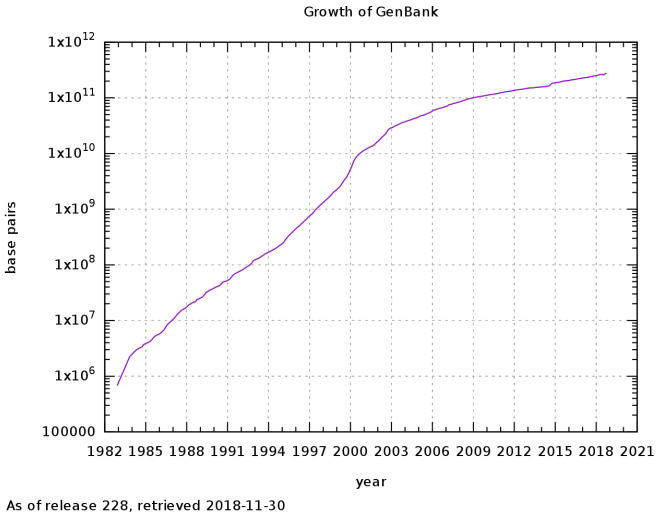


Figura 3: Crescita di GenBank

Tutti questi dati hanno bisogno di essere processati in modo da ottenere informazioni utili. Come prima cosa è possibile estrarre struttura e posizione dei geni dalle sequenze del genoma [6], inoltre è anche possibile identificare gli elementi regolari [7]. Esistono inoltre approcci per estrarre caratteristiche comuni tra gli RNA noti per la previsione di nuovi geni RNA nelle regioni non annotate dai genomi procariotici e arcaici [8].

Nella proteomica i metodi computazionali vengono usati per la previsione della struttura delle proteine. Le proteine hanno una struttura molto complessa composta da migliaia di atomi e legami, di conseguenza il numero di possibili strutture è estremamente elevato. La previsione della struttura risulta quindi un problema combinatorio molto complesso e richiede una buona ottimizzazione.

[1] In genetica, il termine locus genico (o più semplicemente locus, plurale loci) designa la posizione, stabile, di un gene o di un marcatore genico all'interno di un cromosoma

Attraverso l'applicazione dei metodi computazionali è possibile anche gestire dati sperimentali complessi, il campo più noto dove troviamo applicazioni di questo tipo sono i *microarray* [2] nel quale si cerca di identificare i pattern di espressione, cercare una classificazione e creare reti genetiche.

Passiamo ora a vedere alcuni esempi pratici di applicazione di tecniche di apprendimento automatico in campo bioinformatico e medico.

1.2.1 Apprendimento automatico per la previsione di varianti non codificanti associate a malattie

All'interno di tutte le variazioni genetiche quelle riguardanti le malattie rappresentano una piccolissima minoranza rispetto all'insieme più ampio di variazioni genomiche non deleterie. Questo squilibrio porta a un sbilanciamento tra gli insiemi, sopratutto nelle regioni regolatore non codificanti del genoma umano. Quello che si vorrebbe fare è usare tecniche di apprendimento automatico per individuare varianti non codificanti associate alle malattie ma la scarsità di osservazioni inficia sulla loro efficacia. Lo stato dell'arte dei metodi basati sull'apprendimento automatico non adottano tecniche specifiche che tengono conto dello sbilanciamento dei dati e questo porta inevitabilmente a una riduzione della sensibilità e della precisione.

In questo contesto gli algoritmi di apprendimento classici come *support vector machine* [9] o *artificial neural networks* [10] non riescono a generalizzare in maniera sufficiente poiché di solito prevedono la classe di minoranza con una precisione e sensibilità molto bassa. Nel campo della previsione di varianti genetiche associate a tratti o malattie, ciò si riduce a prevedere erroneamente la maggior parte delle varianti associate alla malattia come non associate alla malattia stessa, limitando così in modo significativo l'utilità dei metodi di apprendimento supervisionato per la previsione di nuove varianti non codificanti associate alla malattia.

Per affrontare questo problema è stato sviluppato *hyper*SMURF [11] sigla che sta a indicare: *hyper-ensemble of SMOTE Undersampled Random Forest*. Questo metodo adotta strategie di apprendimento che tengono conto dello sbilanciamento, ovvero tecniche di ricampionamento e su un approccio *hyper-ensemble*: simultaneamente la classe di minoranza viene sovracampionata e la classe di maggioranza viene sottocampionata in modo da generare dati di addestramento bilanciati, ciascuno dei quali verrà poi usato per addestrare un insieme di random forest 1.1.1.2. La sua struttura è visibile in Figura 4.

[2] Si tratta di sottili supporti di materiale plastico o vetro su cui si trovano molte migliaia di pozzetti, ciascuno contenente pochi picogrammi (1 pg = 10-12g) di una diversa sonda di DNA a singola elica.

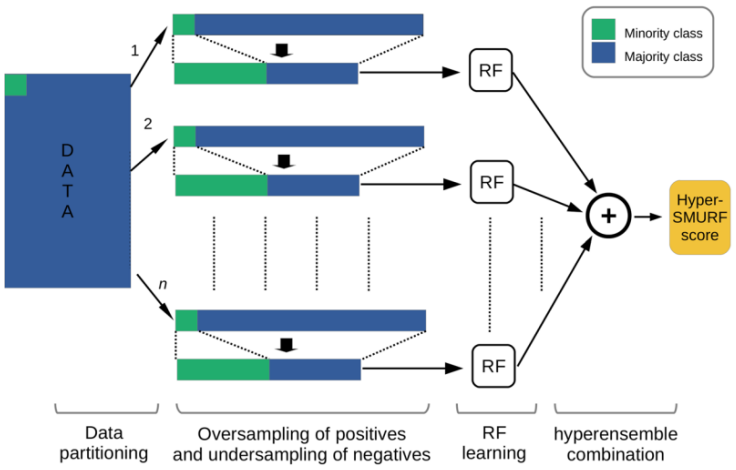


Figura 4: Schema di *hyper*SMURF. Esso divide la classe maggioritaria (rettangoli blu) in n partizioni e per ognuna di esse vengono usate tecniche di sovracampionamento in modo da generare esempi aggiuntivi dalla classe minoritaria (rettangoli verdi). Contemporaneamente un numero simile di esempi viene sottocampionato dalla classe maggioritaria. Successivamente *hyper*SMURF addestra in parallelo n random forest usando i dati bilanciati e infine combina le predizioni degli n insimi secondo un approccio *hyper-ensemble*

.

Successivamente le predizioni dei modelli addestrati sono combinati attraverso un approccio chiamato *hyper-ensemble*, insieme di insiemi, per ottenere una predizione complessiva "concordata". Addestrando n *random forest*, una per ogni insieme di dati, e combinando le loro previsioni facendo la media delle probabilità si ottiene un *hyper-ensemble* poiché ogni dato passato è a sua volta un insieme di alberi decisionali.

Il vantaggio risiede nel fatto di avere molta diversità nei dati di addestramento, inoltre il bilanciamento tra esempi positivi e negativi evitano che si abbia una polarizzazione verso la classe maggioritaria. Tutto ciò rende il modello più capace di generalizzare, mentre l'approccio di *hyper-ensemble* fornisce apprendimento più accurato e previsioni più robuste. Per testare il modello viene utilizzato una 10-*cross-validation* per assicurarsi che le varianti della stessa malattia non si presentino insieme nel *training set*, o insieme di addestramento, e in quello di test, falsando così i risultati. La misura vera e propria è affidata all'AUPRC, ovvero *Area Under the Precision and Recall*, all'AUROC, *Area Under the Receiver Operating Characteristic* e usando la precisione, il richiamo e il punteggio F in funzione della soglia di punteggio.

I risultati sono visibili in figura 5 dove è possibile vedere un confronto tra i vari

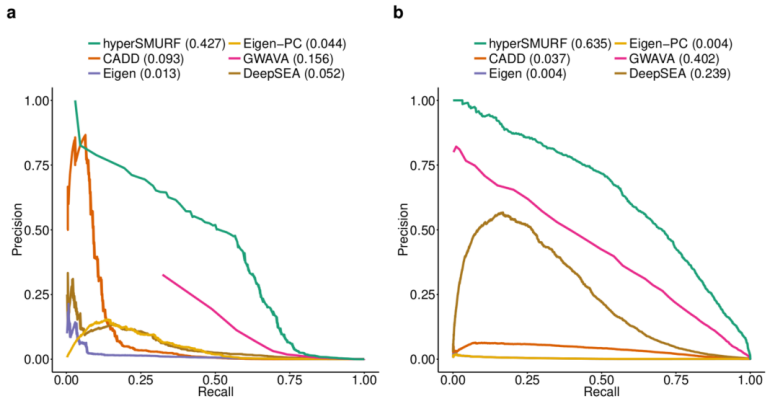


Figura 5: Confronto della curva *precision/recall* tra i vari modelli usando due tipi diversi di dati: a) *Mandelian regulatory mutations*, b) *GWAS regulatory hits*. Il numero tra parentesi rappresenta il valore di AUPRC.

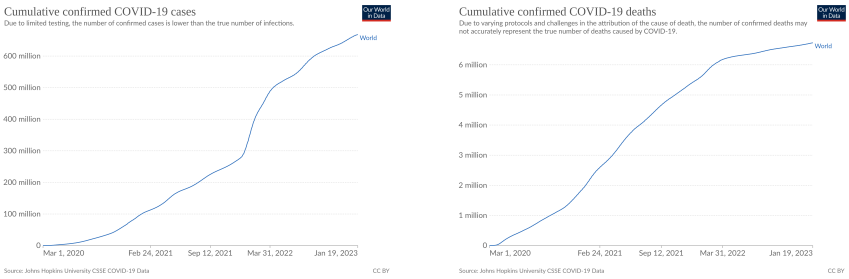
metodi usando la curva *prevision-recall*. Si nota chiaramente come *hyper*SMURF raggiunge migliori risultati rispetto ai metodi classici.

1.2.2 Apprendimento automatico per la previsione del rischio di COVID-19

Negli ultimi anni una grave sindrome respiratoria (SARS-Cov-2) ha colpito il mondo provocando una pandemia mondiale causando 663 milioni di casi accertati e quasi 7 milioni di morti accertati[3].

Da quando il virus ha iniziato a diffondersi tra la popolazione ospedali e pronto soccorsi sono stati invasi da pazienti per i quali era necessario sapere se avessero contratto la malattia o meno ma sopratutto la gravità. Il virus è in grado di causare anomalie nelle radiografie del torace (CXR) ma a causa della sua bassa sensibilità sono necessari ulteriori considerazioni e criteri per riuscire a prevedere il rischio di aver contratto il virus con conseguente gravità. L'obiettivo è quello di riuscire a creare un sistema che riesca a estrarre le variabili ragiologiche, cliniche e di laboratorio più rilevanti che siano in grado di migliorare la previsione del rischio del paziente. Si vogliono inoltre ottenere criteri utilizzabili dai medici nel momento in cui devono decidere il rischio del paziente, intesa come gravità della malattia. In [12] viene illustrato un modello di previsione del rischio per il paziente. Il modello è in grado di selezionare le più importanti variabili cliniche e di laboratorio tenendo anche conto di punteggi radiologici che derivano dalla

[3]https://ourworldindata.org/covid-deaths



(a) (b)

Figura 6: Grafico che mostra a) casi totali nel mondo da quando è scoppiata la pandemia a oggi; b) morti totale nel mondo da quando è scoppiata la pandemia a oggi.

valutazione del CXR [4] da parte dei radiologi e due punteggi di "coinvolgimento polmonare" calcolati usando alcune delle *deep neural network* più performanti per la diagnosi del rischio da COVID-19. Successivamente viene eseguita una fase di imputazione dei dati mancanti causata dall'integrazione di più fonti di dati. Le tecniche usate sono state la *Multiple Imputations by Chained Equations* (MICE [13]), sia utilizzando la corrispondenza media predittiva (*micePPM*), sia classificatori random forest (*miceRF*) come modello di imputazione di bas, e *missForest* [14]. In seguito viene costruito un algoritmo robusto che combina Boruta [15] [16] con metodi di selezione delle caratteristiche, o (*features*), basati sulle permutazioni e incorporati nelle Random Forest [17] [18].

Le *feature* selezionate sono usate come input per random forest (RF) 1.1.1.2 e per alberi associativi derivati (AT) [19]. A differenza delle RF che producono un gran numero di regole di difficile comprensione, le AT sono costruite partendo dalle RF ma producendo un numero regole più semplice che può essere facilmente valutato e interpretato dai medici. I risultati ottenuti dai due algoritmi sono comparati a quelli ottenuti dai modelli generalizzati (GLM [20]) eliminando così l'ipotesi di normalità.

	model	AUC (var)	Sensitivity (var)	Specificity	F1-score	Accuracy
missForest	RF	**0.81** (0.00007)	**0.72** (0.00016)	0.76 (0.00006)	**0.62** (0.00009)	**0.74** (0.00006)
	AT	0.67 (0.00013)	0.51 (0.00039)	0.83 (0.00020)	0.53 (0.00028)	0.67 (0.00013)
	GLM	0.80 (0.00001)	0.56 (0.00002)	0.86 (0.00001)	0.62 (0.00002)	0.71 (0.00001)
miceRF	RF	0.79 (0.00011)	0.70 (0.00034)	0.74 (0.00012)	0.60 (0.0002)	0.72 (0.00014)
	AT	0.65 (0.00027)	0.48 (0.00079)	0.82 (0.00022)	0.50 (0.00062)	0.65 (0.00027)
	GLM	0.78 (0.0005)	0.53 (0.00025)	0.85 (0.00004)	0.59 (0.00014)	0.69 (0.00009)

Figura 7: Tabella con le misure di performance calcolate usando i vari algoritmi

[4] *Chest-X-Ray*

Capitolo 2

Dataset

Un insieme di dati, chiamato *dataset* in inglese, è una raccolta di dati in cui ogni riga rappresenta un'osservazione, o istanza, e ogni colonna costituisce un attributo, o caratteristica, dell'istanza.

L'apprendimento automatico ha bisogno di dati per poter addestrare i modelli. L'obiettivo è riuscire a creare modelli che siano in grado di generalizzare e prevedere correttamente al passaggio di nuovi dati. I dati possono influire significativamente sulle prestazioni nelle predizioni, pertanto è importante utilizzare un insieme di dati ben equilibrato, rappresentativo e qualitativo.

Nel mio lavoro ho usato apprendimento supervisionato quindi l'insieme di dati aveva anche un'etichetta. L'obiettivo è addestrare il modello in modo che riesca a estrapolare la relazione che sussiste tra i dati, X, e l'etichetta, y, per far si che, al passaggio di nuove istanze, esso mi restituisca l'etichetta predetta sulla base dei dati passati.

2.1 Dati Multi-Omici

I dati multi-omici sono un insieme dei dati che contengono le variazioni molecolari su più livelli quali: genomica, epigenomica, trascrittomica, proteomica, metaboloma e microbiotica. La disponibilità di questo tipo di dati ha completamente rivoluzionato il campo della medicina e della biologia.

La genomica è il campo che si occupa dell'identificazione dei geni e delle varianti geniche associate a una malattia o in risposta a determinati medicinali.

Il termine epigenomica si riferisce all'identificazione di modificatori di DNA o delle proteine associate al DNA. Le modifiche epigenetiche del genoma possono anche agire come marcatori per sindromi metaboliche, malattie cardiovascolari e disturbi fisiologici. Queste modifiche possono essere specifiche per cellula e tessuto,

è quindi fondamentale identificare le modifiche epigenetiche durante gli stati nativi della malattia.

La trascrittomica esamina i livelli di RNA in tutto il genoma, sia in modo quantitativo che qualitativo, includendo quali trascrittori sono presenti e i livelli della loro espressione. Sebbene solo il 2% del DNA venga tradotto in proteine, quasi l'80% del genoma viene trascritto, tale processo include l'RNA codificante, RNA corto, microRNA, piwi RNA e piccoli RNA nucleari. Oltre a fungere da intermediario tra DNA e proteine, l'RNA ha anche funzioni strutturali e regolatorie. E' stato dimostrato che essi hanno un ruolo nell'infarto miocardico, nella differenziazione adiposa, nel diabete, nella regolazione endocrina, nello sviluppo neuronale e altri [21],

Il termine proteomica indica il campo nel quale si cerca di identificare i livelli, le modifiche e le interazione delle proteine a livello del genoma. Queste modifiche sono coinvolte nella manutenzione della struttura e della funzione cellulare.

Il metaboloma comprende tutti i metaboliti presenti in una cellula, tessuto o organismo, compresi i piccoli molecolari, carboidrati, peptidi, lipidi, nucleotidi e i prodotti catabolici. Rappresenta il prodotto finale della trascrizione genica e consiste sia di molecole di segnalazione che strutturali. La dimensione del metaboloma è molto più piccola rispetto alla dimensione del proteoma e quindi è più facile da studiare.

Infine la microbiomica consiste in tutti i microorganismi di una comunità. I microbi sono ovunque, sono stati trovati sulla pelle umana, sulle superfici mucose e nell'intestino. Il microbiome presente negli esseri umani è molto complesso, per esempio nell'intestino troviamo circa 100 trilioni di batteri. La microbiota è stata trovata coinvolta in diabete, obesità, cancro, colite, malattie cardiache e autismo. Pertanto, la caratterizzazione del microbiome o di un organismo è di grande interesse medico.

Grazie alla ricerca ci sono stati progressi in diversi campi omici e si è capito che la risposta a una domanda non è da ricercare in un solo tipo di dato poiché questi dati si influenzano tra di loro, esempio il microbiome influenza l'espressione genica e proteica che a sua volta influenza il metabloma. È quindi necessario studiare i dati nella loro interezza sfruttando i dati multi-omici per comprendere lo stato nativo e alterato di un organismo attraverso l'analisi dei dati provenienti da diverse fonti omiche.

Esistono diversi database che forniscono dati multi-omici di pazienti, uno di questi è il *The Cancer Genome Atlas*.

2.1.1 The Cancer Genome Atlas (TCGA)

Il Cancer Genome Atlas è il programma di riferimento per la genomica del cancro, ha favorito la caratterizzazione sistematica di diverse alterazioni genomiche alla

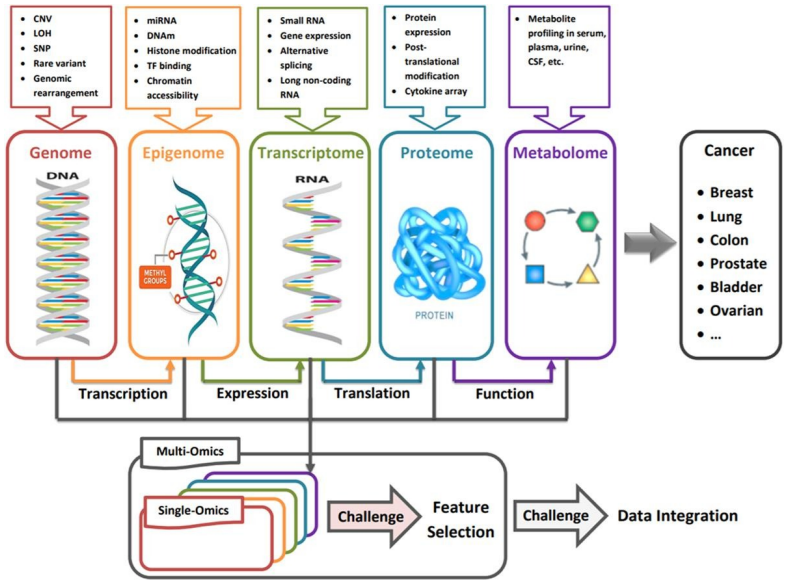


Figura 8: Diagramma delle relazioni tra i vari tipi di dati omici

base dei tumori umano. Attualmente ha caratterizzato oltre 11000 tumori di 33 tipi di cancro. Si tratta di uno sforzo congiunto tra Nation Cancer Institute e il National Human Genome Research Institute e riunisce ricercatori di diversi discipline e istituzioni[22].

I dati utilizzati sono stati presi da TCGA tramite il package curatedTCGAData [1]. Essi sono dati sul carcinoma mammario invasivo (TCGA-BRCA). Questo è un tipo di cancro che si sviluppa all'interno della mammella ed è in grado di diffondersi alle altri parti del corpo attraverso linfonodi o vasi sanguigni. I dati sono stati ottenuti facendo riferimento alla versione del genoma umano hg19, questa versione è stata pubblicata nel 2009 dal progetto internazionale del genoma umano (IGHP) [2] e contiene informazioni sulla sequenza del DNA umano, sequenze dei geni, i siti di regolazione genica e le regioni non codificanti del DNA. Nello specifico, le tipologie di dati utilizzate sono le seguenti:

2.1.1.1 Proteine

Rappresentano i livelli di espressione delle proteine. Le proteine sono le macromolecole che svolgono essenzialmente ogni compito all'interno della cellula. Tra queste, i più importanti sono:

[1] https://bioconductor.org/packages/release/data/experiment/html/curatedTCGAData.html
[2] https://www.genome.gov/human-genome-project

- strutturali: alcune proteine formano la struttura delle membrane cellulari e altre sostengono la forma della cellula,
- catalitiche: molte proteine sono enzimi, che catalizzano reazioni chimiche all'interno della cellula,
- regolatorie: alcune proteine regolano l'espressione genica e la risposta allo stress,
- trasporto: alcune proteine fungono da trasportatori di ioni e molecole attraverso la membrana cellulare,
- immunitarie: alcune proteine fungono da anticorpi, che proteggono la cellula da agenti esterni,
- comunicazione: alcune proteine fungono da messaggeri chimici, che comunicano tra le cellule.

I dati utilizzati durante gli esperimenti sono stati ottenuti con una tecnica nota come RPPA [23] (Reverse Phase Protein Array).

2.1.1.2 mRNA

I dati indicano i livelli di espressione dell'RNA messaggero (noto con l'abbreviazione di mRNA o con il termine più generico di trascritto) è una delle principali molecole in grado di trasportare l'informazione genetica dal nucleo della cellula, dove si trova il DNA, al citoplasma, dove avviene la sintesi proteica. Questi dati sono ottenuti tramite una tecnica chiamata RNA-sequencing [24].

2.1.1.3 miRNA

I dati indicano i livelli di espressione del microRNA (miRNA). Si tratta di piccoli RNA (non codificanti per proteine) che svolgono funzioni di regolazione all'interno della cellula. I livelli di espressione sono ottenuti sempre tramite RNA-sequencing.

2.1.1.4 CNV

Si riferisce al tratto genetico che coinvolge il numero di copie di un particolare gene presente nel genoma di un individuo. Le varianti genetiche, tra cui inserzioni, delezioni e duplicazioni di segmenti di DNA, sono anche collettivamente chiamate varianti del numero di copie. Le varianti del numero di copie rappresentano una significativa proporzione della variazione genetica tra gli individui. Chiamato anche CNV. Per ogni gene viene indicato se lo stesso ha subito delezione (i.e. perdita di una o più copie), amplificazione (acquisizione di una o più copie), neutro (nessuna

modifica). Si possono trovare al suo interno i seguenti valori: 0 (neutro), 1 o 2 (amplificazioni), -1 o -2 (delezioni).

2.1.1.5 Etichetta

Sono le etichette binarie da predire. Esse sono state scaricate da un dataset noto come TCGA-CDR [25], curato manualmente per avere dati clinici e di sopravvivenza il più affidabili possibile. Nello specifico, le etichette fanno riferimento ad una misura nota come PFI (Progression Free Interval) dove:

- 1 indica che il paziente ha un nuovo evento tumorale, che sia una progressione della malattia, una recidiva locale, una metastasi a distanza, nuovi tumori primari in tutti i siti o sia morto con il cancro senza nuovo evento tumorale, compresi i casi con un nuovo evento tumorale il cui tipo è N/A,
- 0 altrimenti.

Capitolo 3

Esperimenti

3.1 Preprocessing

3.1.1 Scalare i dati

3.2 Feature selection

Nel machine learning e in statistica, con il termine *feature selection* si intende il processo di selezione di un sottoinsieme di *feature*, anche chiamate caratteristiche o dimensioni rimuovendo *feature* irrilevanti, ridondanti o che producono solo rumore. Questa pratica di solito porta a una migliore capacità di addestramento, accuratezza più elevata, minore costo di computazione e aumento dell'interpretabilità del modello. La feature selection aiuta anche a non incappare nel *curse of dimensionality*.

Negli ultimi anni i dati disponibili per applicazioni di machine learning in ambiti come mining di testo, computer vision e biomedico stanno aumentando esponenzialmente sia in termini di campioni sia in termini di numero di dimensioni. L'enorme numero di feature dei dataset attualmente disponibili porta a diversi svantaggi: rallentamento significativo degli algoritmi di *learning*, peggiorare la performance dei suddetti algoritmi ma anche portare a una difficile interpretazione del modello. Le tecniche di feature selection possono essere classificate in tre famiglie: metodi supervisionati, metodi semi-supervisionati e metodi non supervisionati.

3.2.1 Tecniche univariate

3.2.1.1 Bassa variabilità

3.2.1.2 Mann-Whitney

The Mann-Whitney U test is a nonparametric test of the null hypothesis that the distribution underlying sample x is the same as the distribution underlying sample y. It is often used as a test of difference in location between distributions.

3.2.2 Tecniche multivariate

3.2.2.1 Minimum Redundancy Maximum Relevance: mrmr

3.2.2.2 Boruta

3.2.3 Dimensionalità intrinseca

3.2.3.1 ID_twoNN

3.2.4 Maximal information-based nonparametric exploration (MINE)

3.2.4.1 The maximal information coefficient (MIC)

3.2.5 Spearman

3.3 Feature extraction

Il termine di *feature extraction*, o estrazione delle caratteristiche, si riferisce al processo di trasformazione dei dati grezzi in caratteristiche numeriche che possono essere elaborate preservando le informazioni nel set di dati originale. Produce risultati migliori rispetto all'applicazione dell'apprendimento automatico direttamente ai dati grezzi.

3.3.1 Uniform Manifold Approximation: umap

3.3.2 t-SNE

3.4 Model selection

3.4.1 Tuning degli iperparametri

3.5 Cross Validation

3.6 Metrica di performance

Le metriche di performance sono molto importanti in un processo di *machine learning*. Esse ci indicano se stiamo facendo progressi nella creazione del modello che meglio si adatta ai dati in input. Esistono diverse metriche che possono essere usate a seconda dei problemi cui siamo davanti. Se stiamo trattando un problema di regressione, avente quindi output continuo, dobbiamo calcolare in qualche modo la distanza tra il dato predetto e quello originale; per fare ciò possiamo usare diverse metriche: *Mean absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE),* R^2 *(R-Squared).*

Siccome il problema affrontato è un problema di classificazione entreremo più nel dettaglio in questo argomento. I modelli di classificazione hanno un output discreto quindi abbiamo bisogno di metriche che comparino classi discrete. Le metriche di classificazione valutano le prestazioni di un modello e ti dicono quanto è buona o cattiva la classificazione, ma ognuna di esse la valuta in modo diverso. Esistono diverse metriche:

- accuratezza,
- matrice di confusione,
- precision e recall
- f1-score,
- au-roc.

3.6.0.0.1 Accuratezza L'accuratezza è la metrica più semplice da usare e implementare. Essa non è altro che il numero di predizioni che il modello ha fatto correttamente diviso per il totale di predizioni, moltiplicato per 100 per avere la percentuale.

3.6.0.0.2 Marice di confusione La matrice di confusione non è propriamente una metriche ma è molto utile per definire le altre metriche.

3.6.1 Dati sbilanciati

3.6.2 Area sotto la curva precision-recall

L'area sotto la curva precision-recall è un singolo numero che riassume l'informazione della curva precision-recall a diverse soglie. La curva PR viene sempre più usata nel *machine learning*, nei problemi di classificazione, sopratutto quando si ha a che fare con *datasets* sbilanciati, ovvero dove una classe è molto più frequente dell'altra. (aggiungi riferimento a come è composto il mio dataset). In questi contesti la curva PR è da preferire alla curva ROC

3.7 Risultati

3.8 Tecnologie usate

Capitolo 4

Conclusioni e sviluppi futuri

Bibliografia

[1] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 07 2003.

[2] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA, 1998. Association for Computing Machinery.

[3] Qingyong Wang, Liang-Yong Xia, Hua Chai, and Yun Zhou. Semi-supervised learning with ensemble self-training for cancer classification. pages 796–803, 10 2018.

[4] L. Busoniu, R. Babuska, De Schutter, B., and D. Ernst. *Reinforcement Learning and Dynamic Programming Using Function Approximators (1st ed.).* CRC Press, 2010.

[5] Iqbal H. Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):160, Mar 2021.

[6] Catherine Mathé, Marie-France Sagot, Thomas Schiex, and Pierre Rouzé. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*, 30(19):4103–4117, October 2002.

[7] Stein Aerts, Peter Van Loo, Yves Moreau, and Bart De Moor. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*, 20(12):1974–1976, March 2004.

[8] R J Carter, I Dubchak, and S R Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res*, 29(19):3928–3938, October 2001.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[10] C.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, USA, 1995.

[11] Max Schubach, Matteo Re, Peter N. Robinson, and Giorgio Valentini. Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. *Scientific Reports*, 7(1):2959, Jun 2017.

[12] Elena Casiraghi, Dario Malchiodi, Gabriella Trucco, Marco Frasca, Luca Cappelletti, Tommaso Fontana, Alessandro Andrea Esposito, Emanuele Avola, Alessandro Jachetti, Justin Reese, Alessandro Rizzi, Peter N Robinson, and Giorgio Valentini. Explainable machine learning for early assessment of COVID-19 risk prediction in emergency departments. *IEEE Access*, 8:196299–196325, October 2020.

[13] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[14] Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011.

[15] Miron B. Kursa and Witold R. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010.

[16] Miron B. Kursa, Aleksander Jankowski, and Witold R. Rudnicki. Boruta – a system for feature selection. *Fundamenta Informaticae*, 101:271–285, 2010. 4.

[17] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, April 2010.

[18] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, Jan 2007.

[19] Houtao Deng, George Runger, Eugene Tuv, and Wade Bannister. Cbc: An associative classifier with a small number of rules. *Decision Support Systems*, 59(1):163–170, March 2014. Funding Information: This research was partially supported by ONR grant N00014-09-1-0656 . Copyright: Copyright 2014 Elsevier B.V., All rights reserved.

[20] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[21] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, May 2017.

[22]

[23] Cristian Coarfa, Sandra L Grimm, Kimal Rajapakshe, Dimuthu Perera, Hsin-Yi Lu, Xuan Wang, Kurt R Christensen, Qianxing Mo, Dean P Edwards, and Shixia Huang. Reverse-Phase protein array: Technology, application, data processing, and integration. *J Biomol Tech*, 32(1):15–29, April 2021.

[24] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, January 2009.

[25] Jianfang Liu, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, Andrew D. Cherniack, Albert J. Kovatich, Christopher C. Benz, Douglas A. Levine, Adrian V. Lee, Larsson Omberg, Denise M. Wolf, Craig D. Shriver, Vesteinn Thorsson, Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok-Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah,

Noreen Dhalla, Robert Holt, Steven J.M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, HarshaVardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbro, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C.S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavobilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli

Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliante, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc-Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn Bubley, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean-Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Smallridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, Jolanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bita Esmaeli, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy

Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong-Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne-Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl-Friedrich Becker, Klaus-Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin,

Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne-Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jr, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, Armaz Mariamidze, and Hai Hu. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416.e11, 2018.

[26] Andreas Lindholm, Niklas Wahlström, Fredrik Lindsten, and Thomas B. Schön. *Machine Learning - A First Course for Engineers and Scientists*. Cambridge University Press, 2022.

[27] Carl Kingsford and Steven L. Salzberg. What are decision trees? *Nature Biotechnology*, 26(9):1011–1013, Sep 2008.