*Bioinfo\_Project User Manual*

**1) creazione\_manifest.py**

Lo script si occupa di elaborare i file json e manifest relativi a mRNA e miRNA, in modo tale da produrre i file manifest finali che servono ad effettuare il download dei dati.

Vengono usate le seguenti variabili globali, quindi fare attenzione ai nomi dei file utilizzati nello script e presenti nella cartella.

mrnaJson = '2\_mRNA.json'

mrnaManifest = '2\_mRNA\_manifest.txt'

mirnaJson = '1\_miRNA.json'

mirnaManifest = '1\_miRNA\_manifest.txt'

Output:

nuovo\_manifest\_m.txt

nuovo\_manifest\_mi.txt

new\_label\_mi.txt

new\_label\_m.txt

**2) Download dati tramite il client gdc\_client.exe**

Con i seguenti comandi (eseguiti come amministratore) vengono scaricati i file contenuti in **nuovo\_manifest\_m.txt** e **nuovo\_manifest\_mi.txt.**

gdc-client.exe download -m <path\_assoluto\_cartella\_progetto>\n1\_unione\_maninfest\nuovo\_manifest\_m.txt --dir .\mRNA\

gdc-client.exe download -m <path\_assoluto\_cartella\_progetto>\n1\_unione\_maninfest\nuovo\_manifest\_mi.txt --dir .\miRNA\

N.B.: Verificare la conferma di ”Successfully downloaded: 1160”

**3) Creazione dataset**

* Spostare i file contenenti i label ”new\_label\_mi.txt” e ”new\_label\_m.txt” creati al punto (1) nella cartella ./n2\_creazione\_dataset .
* Eseguire gli script ”move\_files\_miRNA.py” e ”move\_files\_mRNA.py”.
* Eseguire gli script ”create\_dataset\_mi.py” e ”create\_dataset\_m.py”.

Output:

dataset\_miRNA.csv

dataset\_mRNA.csv

**4) Preprocessing dei dataset**

Eseguire lo script dataset\_Preprocessing.py due volte, cambiando di volta in volta l’argomento (il file del dataset) dato allo script.

* dataset\_Preprocessing.py dataset\_miRNA.csv
* dataset\_Preprocessing.py dataset\_mRNA.csv

Output:

scaled\_dataset\_miRNA.csv

scaled\_dataset\_mRNA.csv

**5****) Feature Selection Methods**

Eseguire lo script relativo al metodo di FS desiderato e rispondere alle domande fatte su linea di comando: le risposte date eseguono il metodo con parametri diversi.

a) Complete or reduced list of labels? (R = reduced/C = complete)

b) Which dataset do you want to process?(mi= miRNA/m = mRNA)

c) Do you want to save the dataset containing only the most relevant features? (Y/N)

* 1. Reduced = implementa la FS sul dataset contenente un numero di label minore senza considerare tutte le sottocategorie di tumori.

Complete = implementa la FS sul dataset contenente tutti i tipi di label.

* 1. miRNA = usa il dataset relativo al microRNA.

mRNA = usa il dataset relativo all’ mRNA.

* 1. Y = salva un dataset che contiene solo le feature più rilevanti ottenute dal metodo di FS utilizzato.

N = esegue il metodo di FS senza salvare il dataset.

Output:

<FS>\_dataset.csv

Result\_FS.txt