***Semmelweis Egyetem / Bioinformatikai Tanszék / Verseny 2021***

# Programfájlok

## A kód fájlok a következőképpen néznek ki

* **trainer.py**: Ebben a fájlban az adatok elemzéséhez, trenírozásához kapcsolódó kód van.
* **trainer\_functions.py**: Ebben a fájlban az adatok elemzéséhez, trenírozásához kapcsolódó Python függvények vannak elhelyezve. A trainer.py használja.
* **trainer\_variables.py**: Ebben a fájlban az alap elérhetőségi adatok vannak változókba meghatározva. A trainer.py és a classifier.py is használja.
* **classifier.py**: Ez a fájl végzi a training data (alape esetben: data/train.csv) osztályzását, a trainer.py-al készített statisztikai elemzés alapján.

## Az adat fájlok a következőképpen néznek ki

### Bemeneti adatok

* **data/train.csv, data/train.txt, data/train.xlsx** identikus emlőtumorral kapcsolatos bemeneti adatokat tartalmazó adatfájlok.
* **data/mezo\_info.pdf** az előbbi fájlok mezőit írja le pontos részletességgel.
* Kimeneti adatok (példa)
* **data/train\_statistics.csv** a training adatok alapján készített statisztikai elemzés eredménye.
* **data/train\_export.csv** a training adatok alapján készített statisztikai eredmény az eredeti adatok (train.csv) soraiba illesztve.
* **data/train\_export\_chemo\_0\_hormon\_1.csv** a train\_export.csv változata, ahol csak a hormonterápiával volt szignifikáns összefüggés.
* **data/train\_export\_chemo\_1\_hormon\_0.csv** a train\_export.csv változata, ahol csak az adjuváns kemoterápiával volt szignifikáns összefüggés.
* **data/train\_export\_chemo\_1\_hormon\_1.csv** a train\_export.csv változata, ahol mindkét terápia esetén (együttesen) volt szignifikáns összefüggés.

### Dokumentáció elérhetősége

**docs/documentation.docs** A versenyre készített kód projekt részletes leírása, dokumentálása

# Futtatás

## Hogyan futassuk?

* Első körben szükséges eldönteni, hogy új vagy a training bemeneti adatokkal szeretnénk a programot futtatni.
* Ha a training adatok megfelelőek akkor hagyjuk a **trainer\_variables.py**-ban meghatározott változókat változatlanul.
  + Amennyiben más adatokkal szeretnénk dolgozni, adjuk hozzá az adatfile-t csv formátumban, a rendszerhez.
  + Határozzuk meg a bemeneti adat mappáját a my\_data\_folder = "data" változó értékének megváltoztatásával.
  + Határozzuk meg a bemeneti fájl nevét a my\_data\_filename = "train.csv" változó értékének megváltoztatásával.
* Ha szükséges módosítsuk a kimenti fájlok nevét hasonlóképpen
* Futtasuk a **trainer.py** programkódját konzolból.
* A program elkészíti a fent felsorolt kimenti fájlok közül a **data/train\_statistics.csv csv** fájlt.
* Futtassuk a **classifier.py** programkódját konzolból.
* A program elkészíti a **data/train\_export.csv** fájlt, illetve ennek a fent említett további 3 változatát.

## Szükséges futtatói környezet

* python 3.8.5
* python library - pandas 1.2.4
* python library - scypy 1.6.2

# A feladat megoldás magyarázata

A megoldásomhoz Python nyelvet használtam és a **train.csv** file használatát vettem alapul a források közül. Programom során a következő pipline elvet követtem:

## 1. fájl: trainer / Trainer.py

* Az első python scriptem - **trainer.py –** *pandas* library használatával *dataframe*-be olvassa a train.csv file-t
* Ezt követően adj. kemoterápia vagy hormonterápia alapján sub dataframeket hoztam létre, amelyeken az elemzéseket végeztem.
  + Összesen 8 sub dataframet hoztam létre.
  + A hormon terápia 1/0 adj. kemoterápia 1/0 viszonylatában 4-et
    - kemo\_0\_hormon\_0,
    - kemo\_1\_hormon\_0,
    - kemo\_0\_horom\_1,
    - kemo\_1\_hormon\_1
  + Illetve individuálisan 2-2-t.
    - kemo\_0 és kemo\_1 (hormon függetlenül)
    - hormon\_0 és hormon\_1 (kemo függetlenül)
  + Minden elemzést ezen a 8 különböző dataframe viszonyban néztem.
* Elemeztem az adatokat
  + „**grade**”,
  + „**pathologic\_stage**”,
  + „**grade**” és „**size\_documented**”
    - méret szerinti 5 csoport, értékek közötti keresés
    - (1, 20), (21, 40), (41, 60), (61, 80), (81, 200)
  + „**grade**” és „**age\_at\_diagnosis**”
    - kor szerint 3 csoport, értékek közötti keresés
    - (1, 50), (51, 65), (66, 100)
* Eredményeimet **train\_statistics.csv** file-ban tettem közzé. Itt számos egyéb oszlop eredményét vettem alapul így
  + **PFS\_event** (*pfs\_0* és *pfs\_1* oszlopok a *train\_statistics.csv-ben*, ha az érték az adott értéket vette fel, értéke a db sorok száma)
  + **DSS\_event** (*dss\_0* és *dss\_1* oszlopok a *train\_statistics.csv-ben*, ha az érték az adott értéket vette fel, értéke a db sorok száma)
  + mindegyikhez számoltam áltagos hónapot a **PFS\_time\_months** alapján. Ezek a *pfs\_0\_months\_mean*, *pfs\_1\_months\_mean, dss\_0\_months\_mean*, *dss\_1 \_months\_mean* oszlopokban jelennek meg.
  + Minegyikhez számoltam a hónapok st. devianciáját (szórását) *pfs\_0\_months\_mean\_std*, *pfs\_1\_months\_mean\_std*, *dss\_0\_months\_mean\_std*, *dss\_1 \_months\_mean\_std* oszlopokban jelennek meg.
* Származtattam adatokat az előzőekből így minden esetben kiszámoltam az 1-es érték arányát: (érték 1 sorok) / (érték 1 + érték 0 sorok összege): *pfs\_1\_perc* és *dss\_1\_perc* oszlopok.
* Összehasonlítottam a kemo/hormon terápiában részt vett eseteket a kemo/hormon terápiában nem vett esetekkel *proportion\_val\_1\_pfs* és *proportion\_val\_1\_dss* (chemo\_X1\_hormon\_X2 esetekben az X1-es érték), *proportion\_val\_2\_pfs* és *proportion\_val\_2\_dss* (chemo\_X1\_hormon\_X2 esetekben az X2-es érték, többinél üres). Ezekhez az értékekhez számotlam khi-négyzet próbát (oszlopok *\_chi2\_p* véggel, *python scypy library*) az alábbi példa szerint, ahol a pfs és hormon érték kerül összevetésre.

|  |  |  |
| --- | --- | --- |
|  | pfs\_0 | pfs\_1 |
| horm\_0: | 503 | 285 |
| horm\_1: | 738 | 345 |

* A PFS és DSS esetekben, ahol a p-value értéke 0.05-ös szignifikancia szint alatti jeleztem a *significance\_level* oszlopban, illetve numerikusan a *significance* oszlopban.

## 2. fájl / classifier - classifier.py

* Második python programommal - **classifier.py –** az előbb elkészítette *train\_statistics.csv* alapul véve megyek végig az eredeti adatsoron majd készítek egy összesítést a train\_export.csv fájlba, ahol 4 oszlopot hozzátéve az eredeti adatsorhoz magyarázom, hogy az adatelemzésem szerint szignifkánsan jobb terápia került-e alkalmazásra (*significance, significance\_reason)*, illetve mi ennek az alapja *adjuvant\_chemoherapy\_sign* és *hormone\_therapy\_sign*.
* Eredményeimet a két terápia közös vonatkozásában (*train\_export\_chemo\_1\_hormon\_1.csv*), illetve külön-külön is közzé tettem: *train\_export\_chemo\_0\_hormon\_1.csv* és *train\_export\_chemo\_1\_hormon\_0.csv*

# Eredményeim

Az előbb említettek alapján eredményeim, hogy mit tekintek hatásos terápiának és milyen esetben a train\_export\*.csv fájlokban találhatóak meg.

Bármilyen kérdés felmerülése esetén állok rendelkezésre.

Készítette **Dr. Dul Zoltán**,

a Semmelweis EMK 2. Éves hallgatója, fogorvos

[zoltan.dul@gmail.com](mailto:zoltan.dul@gmail.com)