Математички факултет
Универзитет у Београду

Семинарски рад

Кластер анализа епигенома
Серозни цистаденокарцином јајника

Ментор:
Проф. Др Ненад Митић
Катедра за рачунарство и информатику

Студенти:
Лазар Савић 4/2021
Лана Матић 143/2021
Смер: информатика

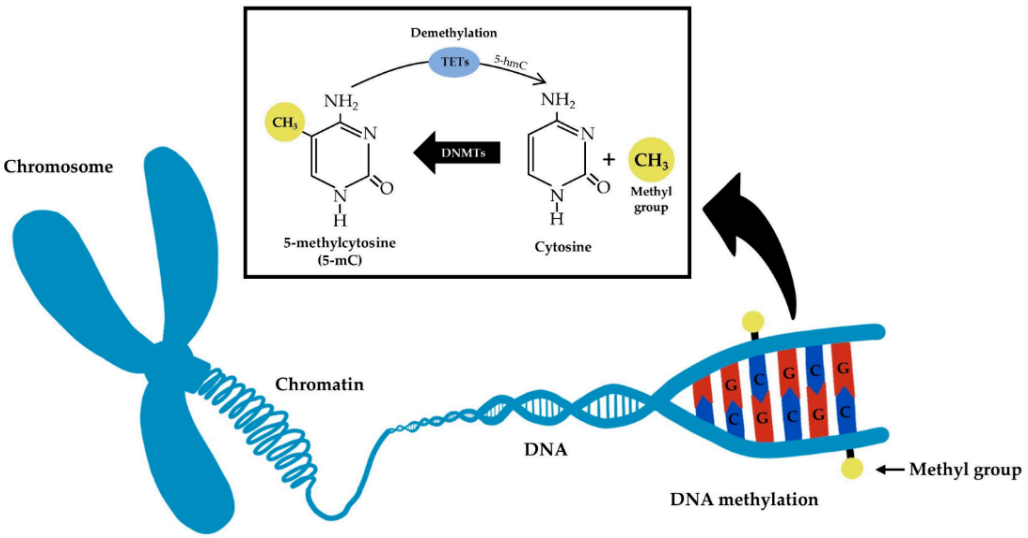
Датум: Фебруар 2025.

САДРЖАЈ

1. Увод

Проучавање и третман различитих врста тумора представљају један од основних задатака савремене медицине. Корен овог проблема и његово разумевање можемо потражити у епигеномици - истраживачком пољу које приказује корелацију између нашег понашања и окружења (као што су исхрана, физичка активност, излагање УВА/УВБ зрачењу, …) и генетске активности. Спољашњи фактори могу проузроковати епигенетке модификације, које укључују или искључују одређене гене, чиме контролишу њихову активност, не утичући на саме гене. Ово представља важан механизам за регулацију биолошких процеса.

Епигенетски процеси утичу на регулацију експресије гена кроз различите механизме, активирајући или деактивирајући гене који су одговорни за синтезу протеина. Један од најтипичнијих епигенетских механизама је *метилација ДНК.* Супституцијом метил (CH3-) групом у одговарајућим регулаторним деловима ДНК (прецизније, на петом угљениковом атому у цитозинском прстену) помоћу ДНК метилтрансферазе, таква структурна промена може утицати као препрека, спречавајући читање гена и тиме онемогућавајући његову даљу експресију. С друге стране, реверзибилан процес уклањања метил групе, процес деметилатације, омогућава активацију гена и његову експресију (*слика 1*) (1).



Слика 1: Упрошћена схематска репрезентација механизама метилације и деметилације ДНК (2)

Серозни цистаденокарцином јајника је малигни облик серозног тумора јајника, што је најчешћи тип епителијалног тумора овог органа, који најчешће има фаталне исходе (3). Разумевање епигенетских промена које прате овај облик рака може дати драгоцене информације за рану дијагнозу, прогнозу и терапију. Применом метода као што је кластер анализа епигенома, истраживачи настоје да идентификују подгрупе пацијената

са сличним епигенетским потписима, што може помоћи у развоју персонализованих терапеутских стратегија.

Овај рад има за циљ да истражи примену кластер анализе епигеномских података серозног цистаденокарцинома јајника, на основу разлике у нивоима метилације (тзв. високоградни и нискоградни облици). Резултати овог истраживања могу допринети бољем разумевању епигенетских профила пацијената који болују од оваквог тумора, што може унапредити стратегије ране дијагнозе и терапије.

2. Подаци - препроцесирање и анализа

Подаци (односно материјал) који анализирамо у овом раду могу се пронаћи на *TCGA Firehose* домену, у датотеци под називом *humanmethylation450-within_bioassay_data_set_function* (доступно на следећем линку). За почетак ћемо приказати како изгледају оригинални подаци и описати њихово значење, а потом ћемо описати како смо те податке обрадили и прилагодили за потребе кластеровања.

Коришћени скуп података представља метилационе профиле CpG острва (тј. локуса) у узорцима туморског ткива јајника, који су добијени *HumanMethylation450 BeadChip* алатом (4). CpG острва су региони у ДНК који садрже високу густину CpG динуклеотида (цитозин - фосфат - гуанин парова), и они су посебно су важна за регулацију експресије гена, јер су често присутна у промоторским секвенцама гена. Хиперметилација ових острва доводи до потискивања експресије гена, јер се транскрипциони фактори не могу везати за промотор. Овај механизам је често одговоран за супресију активности гена који спречавају развој тумора.

Табела 1 илуструје како изгледају необрађени преузети подаци.

Composite Element REF	**Beta_value**	**Gene_Symbol**	**Chromosome**	**Genomic_Coordinate**	**…**
cg00000029	0.1622	RBL2	16	53468112	…
cg00000108	NaN	C3orf35	3	37459206	…
cg00000109	NaN	FNDC3B	3	171916037	…
cg00000165	0.0775	NaN	1	91194674	…
cg00000236	0.8874	VDAC3	8	42263294	…

Табела 1: Приказ сирових преузетих података

Сваки ред ове табеле представља CpG локацију, којој су придружене бета вредности и метаподаци о самој локацији. Структура табеле је таква да се од друге до последње колоне подаци организују у понављајуће блокове од по четири колоне: Beta_value, Gene_Symbol, Chromosome и Genomic_Coordinate. Сваки такав блок представља информације за један биолошки узорак (пацијента), који се идентификује преко Hybridization REF вредности. Укупно има 10 таквих блокова, што значи да подаци обухватају 10 узорака пацијената. Иако су колоне Gene_Symbol, Chromosome и Genomic_Coordinate унутар сваког блока исте, оне се формално понављају у сваком блоку, јер сваки блок стоји као засебна целина у скупу података.

- *Composite Element REF* - јединствени идентификатор CpG локуса
- *Beta_value* - вредност бета метилације у распону [0, 1], где 0 означава потпуно неметиловану, 1 потпуно метиловану, а остале вредности делимично метиловану CpG локацију
- *Gene_symbol* - име гена повезаног са CpG локусом
- *Chromosome* - хромозом на ком се налази CpG локус

- *Genomic_Coordinate* - геномска позиција CpG локације у односу на референтни геном.

У једној врсти, вредности поља *Gene_symbol, Chromosome* и *Genomic_Coordinate* се поклапају, једино се мењају бета вредности (разлог за ово је што се испитују исте позиције у геному за различите пацијенте, који имају различит степен метилације).

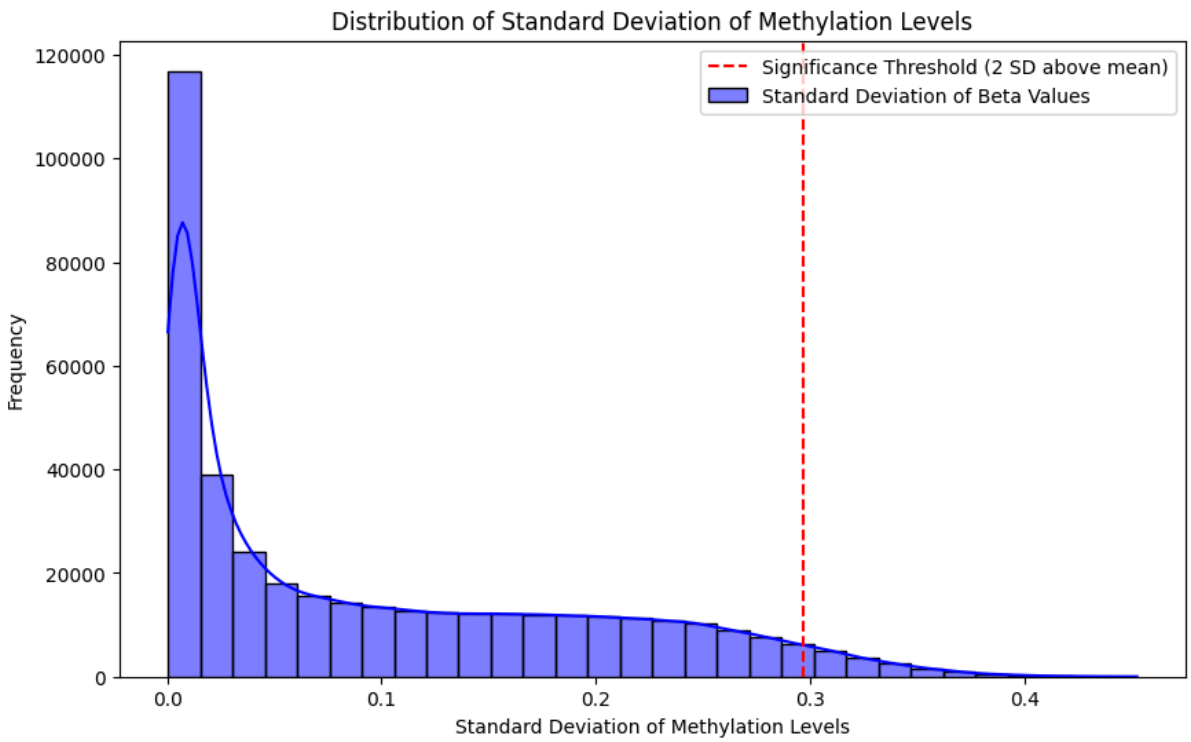
Суштину нашег разматрања представљају бета вредности, јер се на основу њих могу уочити разлике метилационих профила здравих и оболелих особа. То је заправо и мотивација за кластеровање ових података - оно омогућава сегрегацију пацијената на различите подтипове ове болести, што за последицу има раније дијагностиковање и успешније сузбијање даљег напредовања тумора.

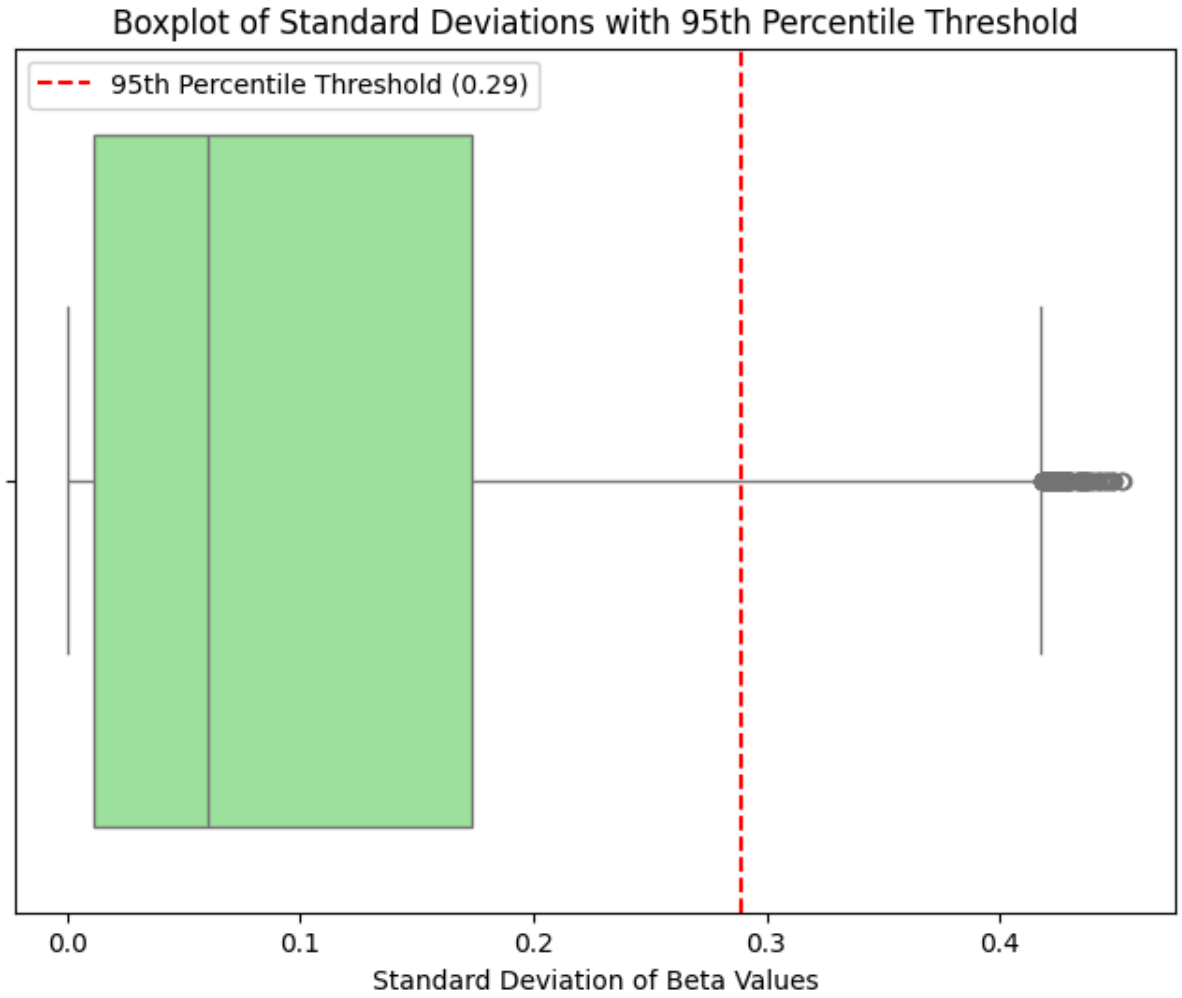
Пре него што пређемо на само кластеровање, податке је неопходно припремити, тј. препроцесирати, затим визуелизовати и упознати се са њиховом природом и значењем. За почетак, испитано је присуство недостајућих вредности и установљено је да одређен број поља у колонама *Beta_value* (а и у осталим колонама) нема вредност. За сваку *Beta_value* колону, испитали смо колико редова садржи NaN вредности у тој колони. Приметили смо да су ти бројеви приближне вредности (око 89730), односно да су у највећем броју случајева непознате све *Beta_value* вредности свих 10 узорака пацијената. У занемарљиво малом броју случајева, непознате су *Beta_value* вредности у само неким од узорака пацијената. Иако постоји велики број метода за импутацију недостајућих вредности, ми смо ипак одлучили да редове којима недостају све (или неке) бета вредности у потпуности избацимо. Разлог за овакав избор је жеља за употребом искључиво реалних (а не вештачки израчунатих, потенцијално пристрасних) података, како би добијени резултати кластеровања били што веродостојнији. Све бета вредности су већ нормализоване, тј. налазе се у дозвољеном интервалу [0, 1]. Све колоне *Gene_symbol, Chromosome* и *Genomic_Coordinate* су идентичне, те су избрисана сва појављивања ових колона (сем првих).

Следећи корак је упознавање са дистрибуцијом података. За почетак, морамо одредити *праг значајног нивоа метилације.* Одређени CpG локуси могу имати значајну *варијацију у метилацији* између различитих узорака (нпр. здраво ткиво и туморско ткиво). Због тога је важно идентификовати CpG регионе који показују **значајну** промену метилације, која нам помаже у:

- Разликовању стабилних CpG региона (оних који не мењају метилацију) од динамичних CpG региона.
- Идентификацији CpG места која су епигенетски маркери рака.
- Дефинисању CpG региона који су **диференцијално метиловани** (DMRs), што значи да показују статистички значајне разлике у метилацији између група.

Најчешћи начин одређивања значајног нивоа метилације укључује рачунање стандардне девијације. CpG локуси са високим стандардним одступањем показују велику варијацију метилације међу узорцима, што их чини биолошки значајним. За сваки CpG локус израчуната је стандарна девијација 10 бета вредности од 10 испитиваних узорака, чиме је добијен узорак стандардних девијација, који је визуелизован наредним хистограмом и кутијастим дијаграмом (*слика 2*).





Слика 2: Приказ расподела стандардних девијација бета вредности CpG локуса

За праг значајног нивоа метилације (означен црвеном испрекиданом линијом на хистограму) узета је вредност од две стандардне девијације изнад математичког очекивања. Приближна вредност прага добија се и када се израчуна 95. перцентил. Та вредност приказана је црвеном испрекиданом линијом на кутијастом дијаграму. Будући да смо на два начина добили сличне вредности прага, израчунали смо аритметичку средину ове две вредности и добијени резултат у наставку користили као праг значајног нивоа метилације. Конкретне вредности приказане су у табели 2.

	Вредност прага	**Број значајних метилационих вредности**
m + 2σ	0.2969	16622
95. перцентил	0.2893	19779
Просек	0.2931	18120

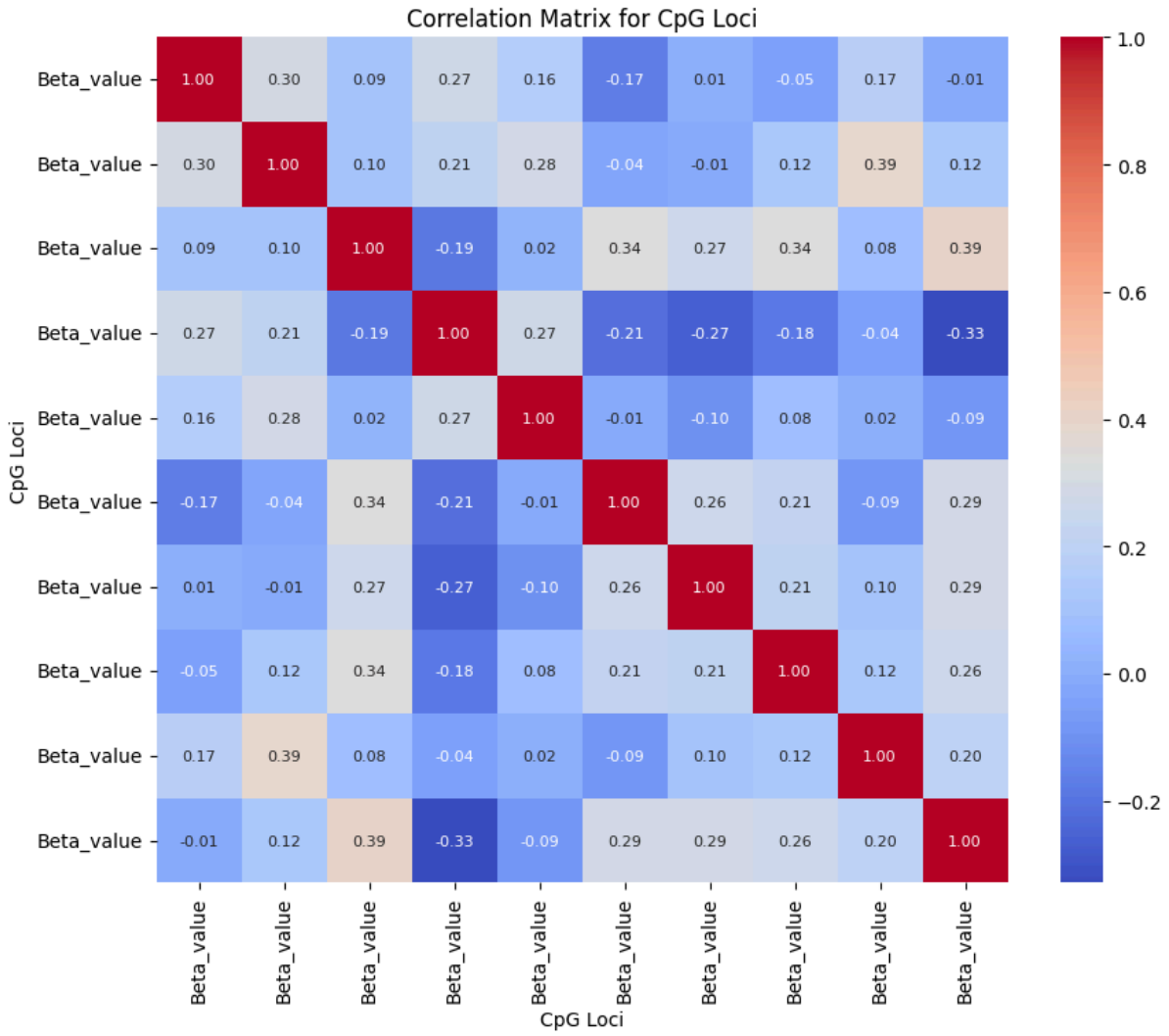
Табела 2: Вредности прага значајног нивоа метилације

Из скупа података избачени су сви редови (односно сви CpG локуси) који не показују значајну промену у нивоу метилације (то су сви редови чија је стандардна девијација бета вредности 10 узорака мања од узете вредности прага). Избачене су све колоне које не носе информацију о бета вредностима, будући да оне нису од значаја за проучавање даљих карактеристика наших података.

За трансформисане податке можемо израчунати матрицу корелације. Корелација између CpG локуса нам омогућава да разумемо потенцијалне метилационе обрасце у геному. Другим речима, желимо да анализирамо да ли се одређени CpG локуси понашају слично у различитим узорцима. Ово је важно из неколико разлога:

- Висока корелација CpG локуса може значити да они припадају истом епигенетском модулу, односно да су део истог регулаторног механизма или да имају заједничку биолошку улогу
- Ниска корелација CpG локуса може указивати на епигенетске дисрупције у канцеру, што указује на CpG локусе који разликују болесне и здраве узорке.

Наредна топлотна мапа приказује матрицу корелације CpG локуса, где је свака ћелија матрице вредност Пирсонове корелације између два CpG региона (*слика 3*).

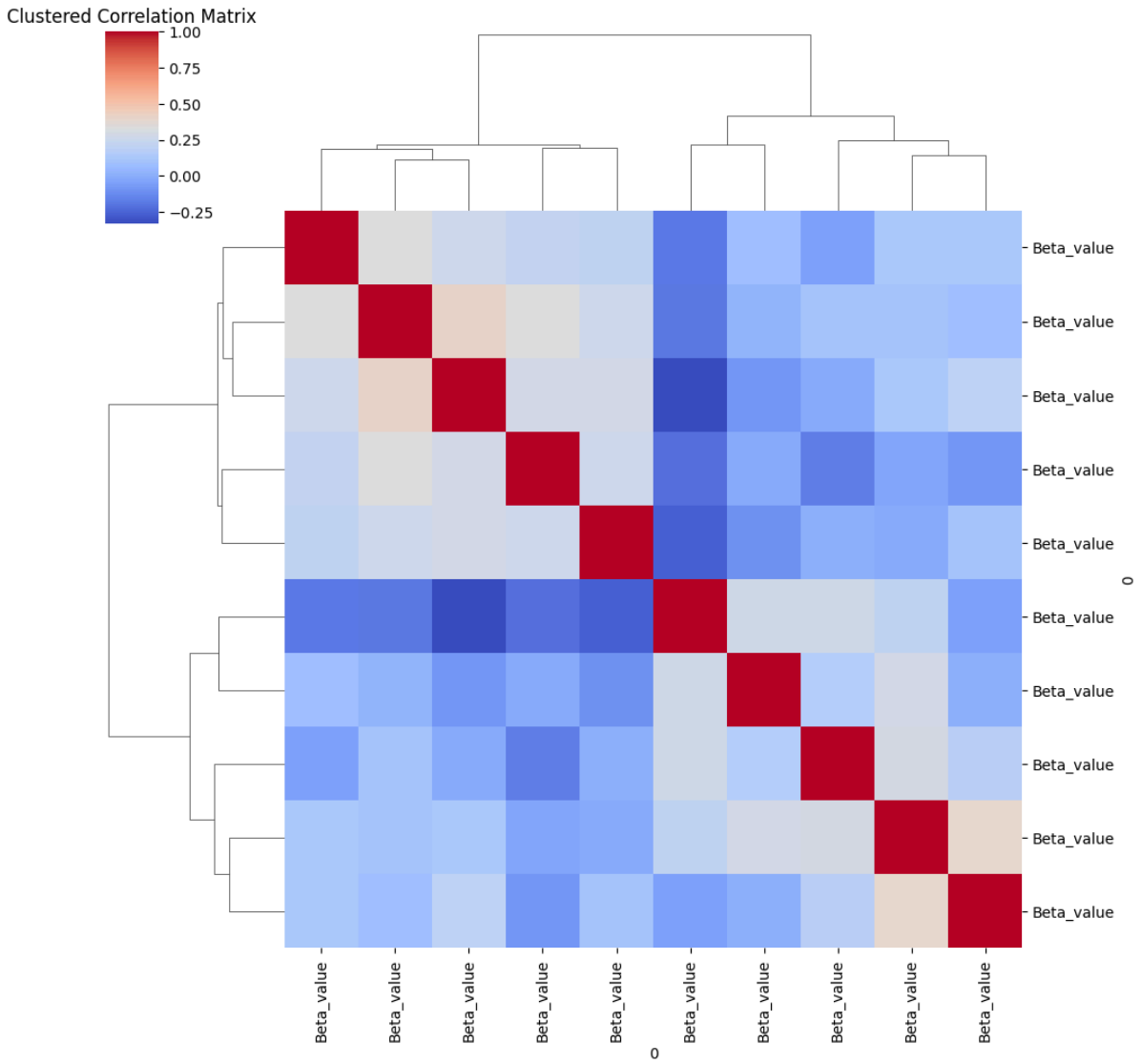


Слика 3: Топлотна мапа матрице корелација CpG локуса

На мапи се уочавају **три различите области**:

1. **Црвена област – високо позитивна корелација** (блиска 1)
    - CpG региони у овој области имају сличне обрасце метилације.
2. **Плава област – негативна корелација** (блиска -1)
    - CpG региони у овој области показују супротне обрасце метилације – када је један регион високо метилован, други је слабо метилован.
    - Ово може указивати на епигенетску регулацију гена, где CpG региони делују као *"метилацијски прекидачи"*, утичући на активност гена.
3. **Беле/светле/неутралне области – нема значајне корелације** (око 0)
    - CpG региони у овој области немају јасну међусобну повезаност у обрасцима метилације.

Како бисмо још јасније увидели корелисаност CpG локуса, извршићемо хијерархијско кластеровање (*слика 4*). Тиме идентификујемо кластере CpG локуса који међусобно корелирају, а такви обрасци су значајни из већ поменутих разлога.



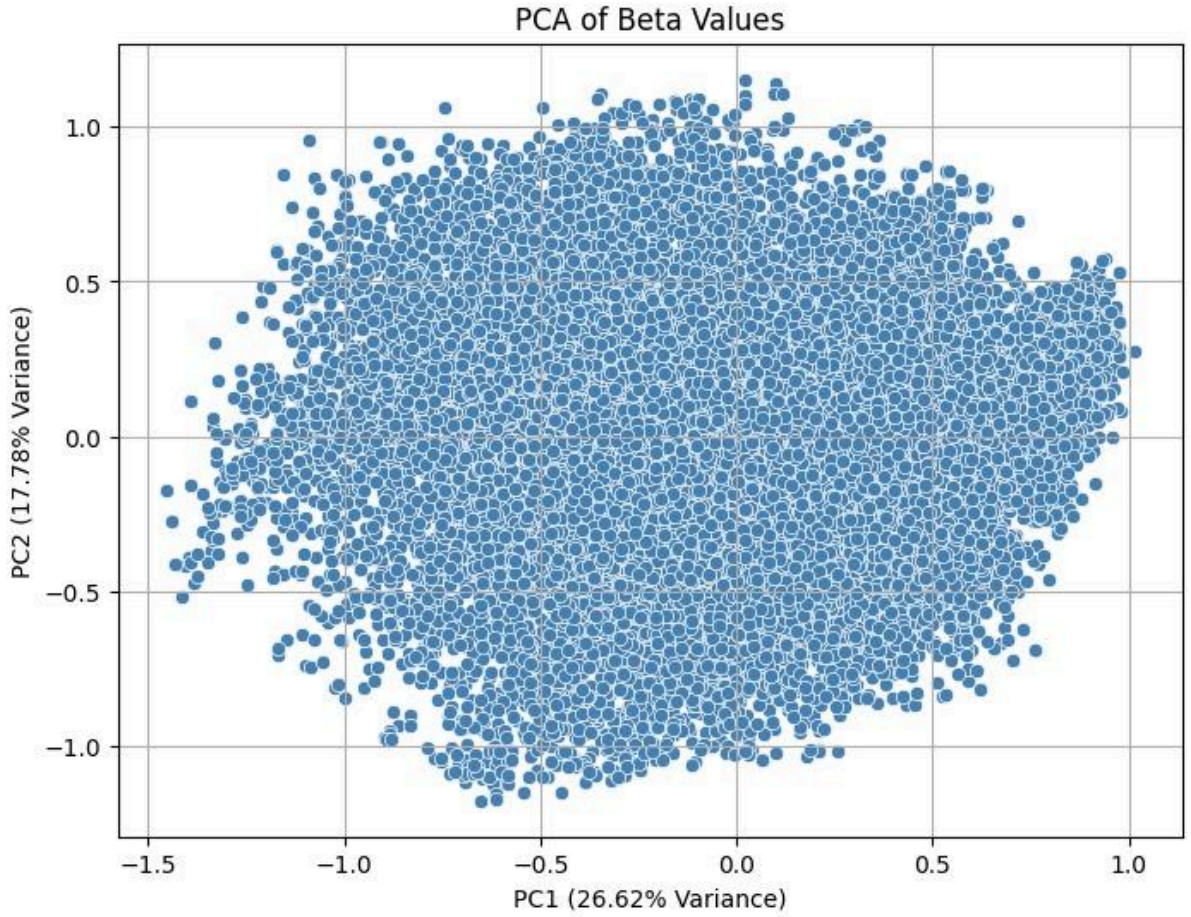
Слика 4: Топлотна мапа са извршеним хијерархијским кластеровањем

Коришћење ових информација у комбинацији са кластеровањем може омогућити прецизније идентификовање епигенетских потписа болести. Оно што се на овој топлотној мапи јасно уочава јесте да су високе вредности корелације ограничене углавном на елементе дуж дијагонале, што је и очекивано, јер сваки CpG локус савршено корелише сам са собом. Иако CpG локуси у топлотној мапи нису били сортирани по геномским координатама, приметили смо да неке групе CpG локуса показују умерено позитивну или негативну корелацију (нпр. 0.39, 0.34, -0.33). Ови обрасци сугеришу да одређени CpG локуси имају сличну епигенетску регулацију, без обзира на то да ли су геномски суседни. Због тога би било корисно испитати метилацијске обрасце не само појединачних CpG локуса, већ целих региона, и то на начин који поштује геномску организацију. Зато смо увели клизне прозоре (поглавље 3.2) - јединице у којима се CpG локуси анализирају заједно, у контексту своје геномске позиције.

На овај начин смо повезали информацију о метилацијској сличности (добијену из топлотне мапе) са геномском контекстуалношћу (обрађеном кроз прозоре), што је омогућило прецизнију анализу и касније кластеровање.

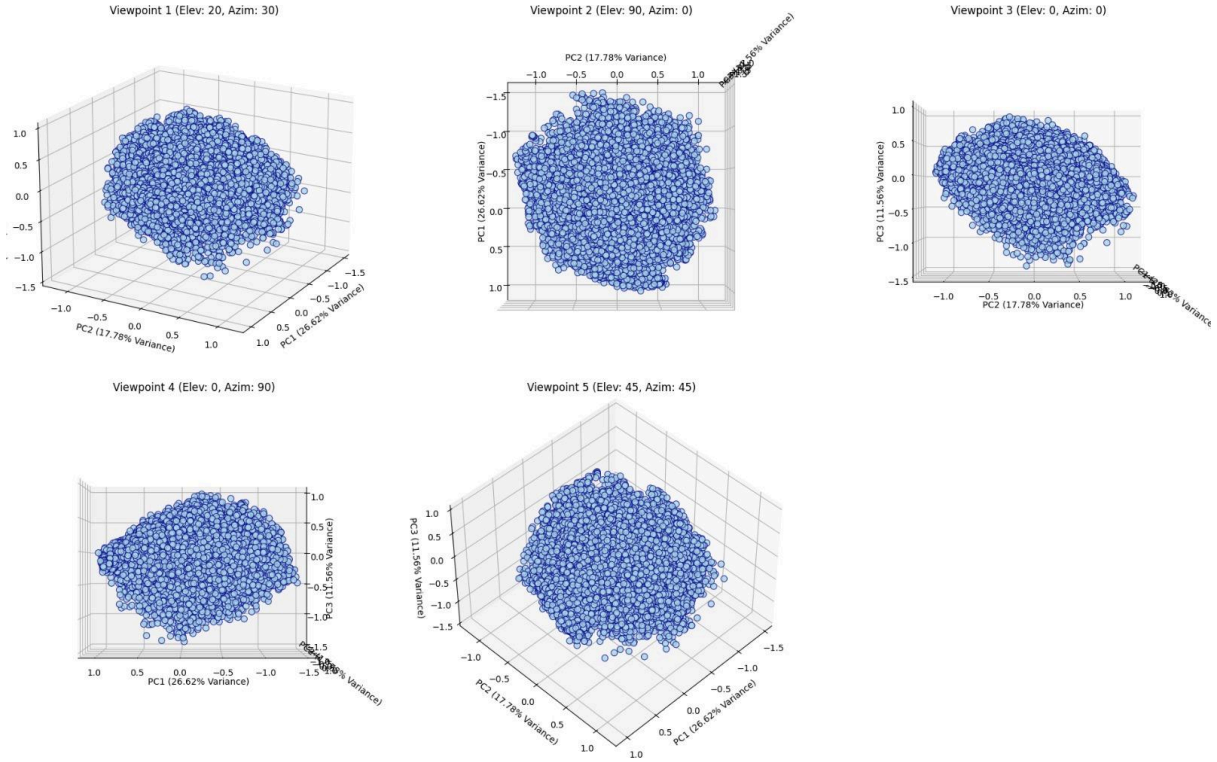
Пре него што почнемо са кластеровањем података применом различитих алгоритама, желимо да стекнемо интуитивну слику о расподели и унутрашњој структури метилацијских података, применом анализе главних компоненти (PCA – Principal Component Analysis). Ова техника смањује димензионалност података, задржавајући при томе што је више могуће укупне варијансе у скупу.

У нашем случају, β-вредности за CpG локусе имају велику димензију (свака тачка представља CpG локус описан вредностима у више узорака). Да бисмо визуелно анализирали структуру података, смањили смо димензију са 10 на 2 (*слика 5*), и на 3 (*слика 6*).



Слика 5: Анализа главних компоненти у 2 димензије

Дијаграм приказује пројекцију података на две главне компоненте, које заједно објашњавају око 44.4% укупне варијансе. Подаци су равномерно распоређени, без јасно одвојених група. Ово може указивати на то да метилацијски обрасци CpG локуса немају јаку унутрашњу кластерску структуру, бар не у прве две компоненте.



Слика 6: Анализа главних компоненти у 3 димензије

Сада су подаци приказани у три главне компоненте, и заједно оне објашњавају приближно 56% варијансе у подацима. Приказано је више перспектива, како би се добио бољи увид у укупну распршеност података у тродимензионалном простору.

Слично као у 2D случају, подаци не показују јасне и компактне кластере, већ чине прилично континуирану и густу расподелу. Ово сугерише да опажања добијена применом анализе главних компоненти (PCA) на две и три главне компоненте није открила јасно одвојене и компактне кластере, већ континуирану и дифузну структуру. Ови резултати указују да CpG локуси, бар у нижим димензионалностима, не показују јаку кластерску организацију. Иако визуелно не уочавамо јасне кластере, то не искључује постојање скривених структура у вишедимензионалном простору. Управо због тога ћемо у наставку применити више алгоритама кластеровања како бисмо омогућили боље разумевање потенцијалних шаблона у метилацијским подацима.

3. Кластеровање

Како бисмо анализирали обрасце метилације CpG региона у серозном цистаденокарциному јајника, примењујемо кластер анализу.

Спровешћемо две различите врсте кластеровања - кластеровање CpG локуса (*поглаље 3.1*), где сваки локус представља вектор од 10 бета вредности (за 10 различитих узорака), и кластеровање региона добијених клизним прозором (*поглавље 3.2*), где се као улаз користи једна агрегатна вредност (средња корелација CpG локуса у прозору). На овај начин се могу открити биолошки значајни обрасци и потенцијални епигенетски биомаркери.

Све анализе и кластеровања изведене у оквиру овог рада реализоване су у програмском језику Python (неопходна верзија >= 3.10) у оквиру окружења Jupyter Notebook, користећи библиотеке: pandas, numpy, scikit-learn, scipy, seaborn, matplotlib, kneed, collections, mpl_toolkits.

Алгоритме за кластеровање које смо користили у овом раду су: K-Means кластеровање, DBSCAN, хијерархијско кластеровање, Gaussian Mixture Model (GMM) и спектрално кластеровање. Метрике коришћене за евалуацију кластеровања су:

- **Силуета коефицијент** (5) - мери квалитет кластеровања анализирајући колико је свака тачка близу сопственом кластеру у поређењу са другим кластерима.

  $$S = \frac{b - a}{max(a,b)}$$

  где је,

  - a - просечно растојање тачке од других тачака у истом кластеру
  - b - просечно растојање тачке од најближег суседног кластера.

  Вредности ближе **1** су боље (добро кластеровање).
  Вредности ближе **0** указују на слабо дефинисане кластере.

- **Калински-Харабаш индекс** (6) - израчунава однос компактности и раздвојености кластера.

  $$CH = \frac{\sum_{i=1}^{k} n_i \left|\left| c_i - c \right|\right|^2}{\sum_{i=1}^{k} \sum_{x \in C_i} \left|\left| x - c_i \right|\right|^2}$$

  где је,

  - c_i - центроид кластера
  - c - центроид читавог скупа

  Веће вредности указују на боље кластере.

Иако појам „центроид“ долази из K-Means алгоритма, његова дефиниција остаје валидна и за друге методе кластеровања, јер се c_i у том случају дефинише као просек тачака (у нашем случају, вектора бета вредности) у кластеру n_i који је добијен том

методом кластеровања. Према томе, Калински - Харабаш индекс остаје применљив и омогућава упоредиву евалуацију различитих алгоритама кластеровања.

- **Дејвис-Болдин индекс** (7) - мери однос компактности и разлике између кластера.

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j} \frac{s_i - s_j}{d(c_i, c_j)}$$

где је,

- s_i - просечно растојање између тачака у кластеру i
- $d(c_i, c_j)$ - удаљеност између два центроида.

Мање вредности су боље јер означавају јасно раздвојене кластере.

Поред ових метрика, за GMM алгоритам коришћене су и специфичне мере као што су Log-Likelihood, AIC (Akaike Information Criterion) и BIC (Bayesian Information Criterion). Детаљније објашњење ових критеријума биће представљено у оквиру прве обраде самог алгоритма (*поглавље 3.1.4*).
Евалуација кластера кроз више метрика помаже у одабиру најбоље методе за анализу CpG метилације.

3.1. CpG локуси

У овом приступу, кластерујемо **појединачне CpG локусе** на основу њихових β-вредности у различитим узорцима. Циљ овог приступа је да се идентификују **CpG региони који имају сличне обрасце метилације**, што може указивати на заједничке регулаторне механизме или улогу у патогенези тумора.

- CpG локуси са сличним метилацијским обрасцима могу припадати **истим регулаторним јединицама** (нпр. промотори, енхансери).
- Овај метод може помоћи у идентификацији CpG локуса који **разликују здраве узорке од туморских**.
- Груписање CpG локуса по сличности може открити **епигенетске потписе специфичне за болест**.

3.1.1. KMeans

Алгоритам K-Means кластеровања је коришћен за анализу CpG локуса, користећи вредности β-метилације као улазне податке.
Користећи параметарску претрагу, *ParametarGrid* из библиотеке *scikit-learn*, испитујемо број кластера, *k,* у опсегу [2, 14], како бисмо пронашли оптималан број кластера у подацима.

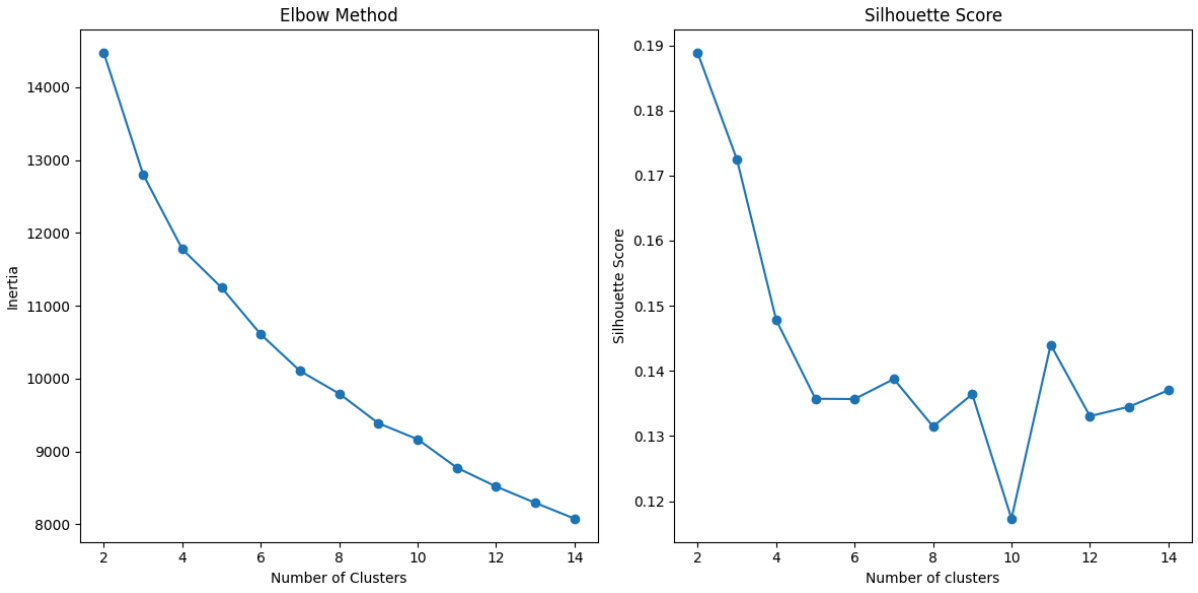
Добијени резултати кластеровања за различите вредности k:

k	Силуета коефицијент	Калински-Харабаш индекс	Дејвис-Болдин индекс
2	**0.1888**	**4175.5390**	2.0398
3	0.1724	3538.9779	1.9800
4	0.1478	3090.4343	1.8727
5	0.1357	2638.9922	1.9524
6	0.1357	2457.9960	1.8783
7	0.1388	2299.5870	1.9231
8	0.1315	2116.3896	1.9750
9	0.1364	2030.0288	1.9322
10	0.1173	1895.7746	1.9372
11	0.1440	1862.5831	1.8400
12	0.1331	1794.3620	1.8216
13	0.1345	1730.1567	1.7905
14	0.1370	1677.1092	**1.7309**

Табела 3: Резултати KMeans кластеровања за различите вредности броја кластера, k

За бољу визуализацију резултата, приказани су графици (*слика 7*) који помажу у одређивању оптималног броја кластера:

- *Лакат метод* (енг. *elbow*) - приказује вредности инерције (суму квадрата удаљености узорака од најближег центра кластера) у зависности од броја кластера. Оштра промена (elbow point) указује на оптималан број кластера.
- *Силуета коефицијент*



Слика 7: Графици вредности метрика добијених KMeans кластеровањем

Додатно, примењен је *KneeLocator* како би се прецизније идентификовала оштра промена (elbow point). Добијени оптимални број кластера применом ове методе је 5.

На основу различитих метрика, добијени су следећи закључци:

- **Силуета коефицијент и Калински-Харабаш индекс** указују да је **k = 2** најбољи избор.
- **Дејвис-Болдин индекс и Elbow метода** сугеришу да је оптималан број кластера **k = 4–5** или можда више.
- Ниске вредности силуета коефицијента и постепено смањење инерције указују да скуп података можда нема јасно дефинисане кластере.
- Све метрике показују ниске резултате, што може значити да су кластери преклапајући или да подаци уопште немају природну кластерску структуру.

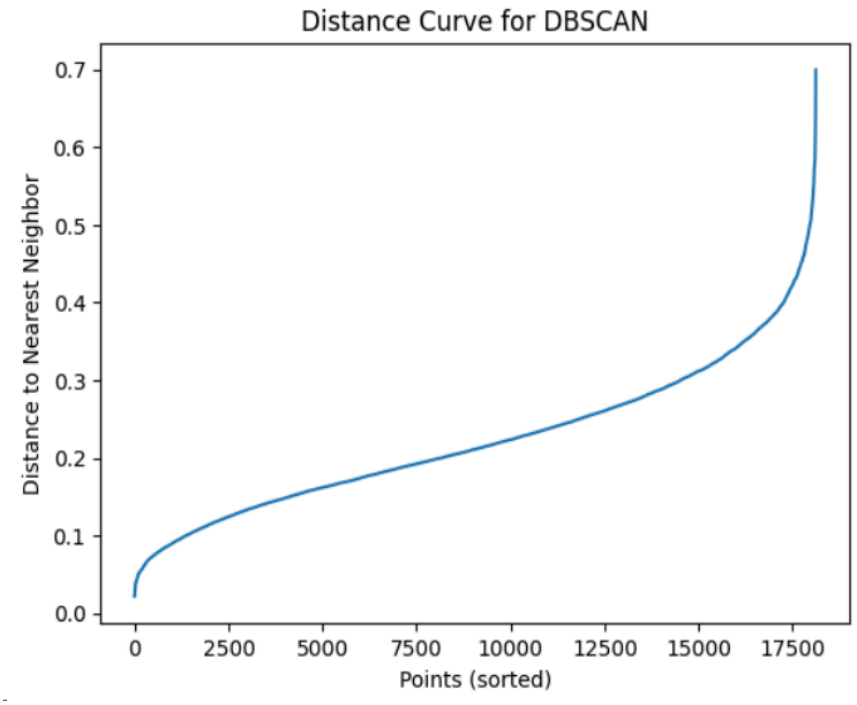
На основу свих анализа, можемо закључити да K-Means алгоритам није одговарајући за овај задатак кластеровања**.**

3.1.2. DBSCAN

Да бисмо кластеровали бета вредности узорака употребом DBSCAN алгоритма, неопходно је да одредимо оптималне вредност два наредна параметра:

- *eps* - максимална удаљеност две тачке да би биле сматране као чланови истог кластера.
- *min_samples* - минималан број тачака који је потребан да се формира кластер.

У сврху одређивања оптималне вредности првог параметра, употребљен је алат *NearestNeighbors* из библиотеке *scikit-learn.* Он омогућава рачунање удаљености најближег суседа за сваку тачку у подацима, које се погодно може визуелизовати употребом *криве удаљености.* Уколико ова растојања сортирамо у растућем редоследу, можемо да добијемо график који приказује како се ова растојања мењају по редоследу. Карактеристичан је моменат наглог раста ове криве, који одговара преласку између густих области тачака (кластера) и области које су “разређене”, па је од суштинске важности добро проценити када се он дешава. На *слици 8* приказана је крива удаљености добијена над подацима које желимо да кластерујемо.



Слика 8: Крива удаљености

Уочава се нагли скок за епсилон вредност 0.4, те је та вредност коришћена у алгоритму. За минималан број елемената у кластеру (вредност *min_samples*) узето је 25.

Са оваквим вредностима параметара, оптималан број кластера износи 3. Међутим, метрике које испитују квалитет добијеног резултата показују да су пронађени кластери лоши (табела 5). Ниске вредности силуета коефицијента и високе вредности Дејвис - Булдин индекса указују на то да обе кластери нису јасно раздвојени, тј. да нису смислени. С тим у вези, закључује се да DBSCAN није адекватан алгоритам кластеровања за коришћене податке.

3.1.3. Хијерархијско кластеровање

Хијерархијско кластеровање је примењено на вредности β-метилације CpG локуса користећи **Ward-ову методу** за израчунавање удаљености између кластера. Ова метода минимизује укупну варијансу унутар кластера.
Користећи параметарску претрагу, *ParametarGrid* из библиотеке *scikit-learn*, испитујемо број кластера, *k,* у опсегу [2, 14], како бисмо пронашли оптималан број кластера у подацима.

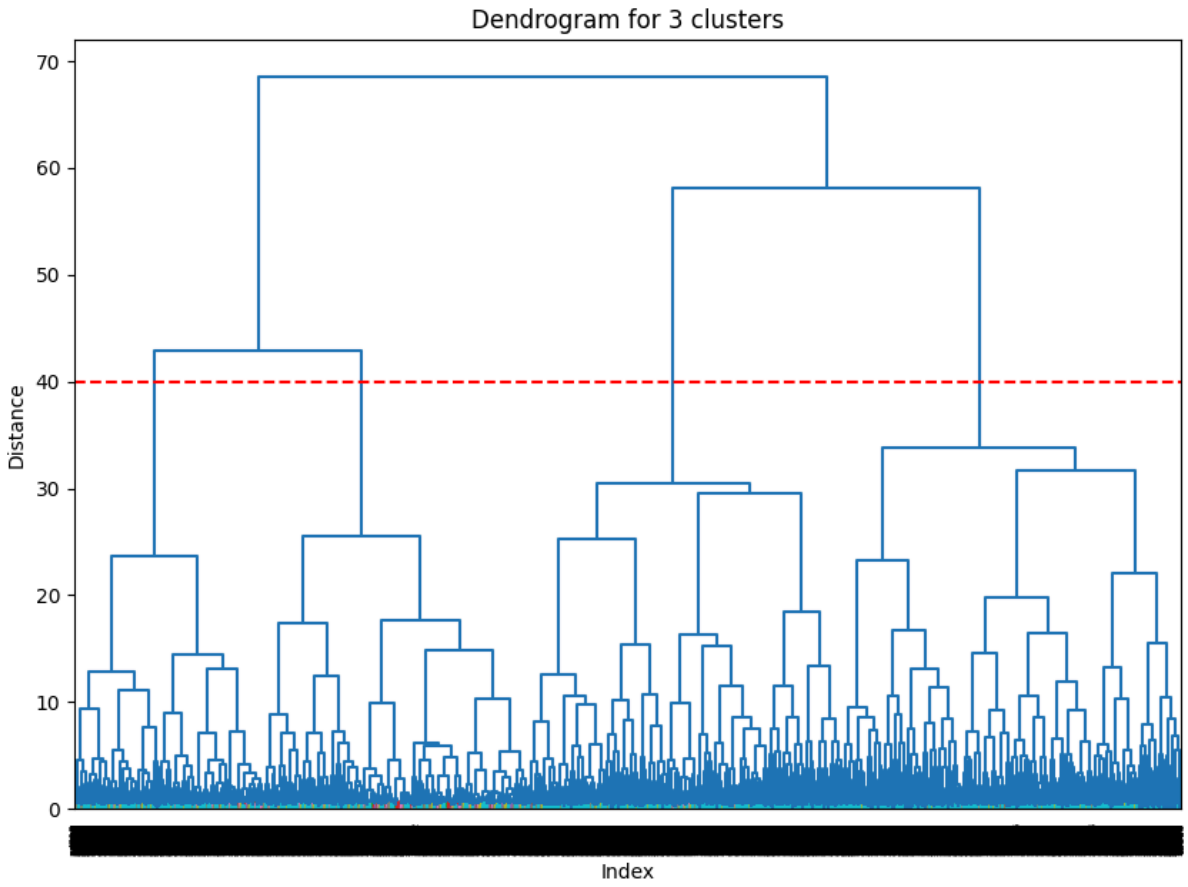
Добијени резултати кластеровања за различите вредности k:

k	Силуета коефицијент	Калински-Харабаш индекс	Дејвис-Болдин индекс
2	0.1215	**2754.2241**	2.3817
3	**0.1357**	2655.8064	2.2801

4	0.0938	2328.9707	2.1463
5	0.0960	2039.6526	2.3168
6	0.1003	1856.0994	2.1328
7	0.1072	1734.3804	2.3455
8	0.1127	1649.5181	2.1892
9	0.0891	1557.7829	2.1681
10	0.0933	1490.4543	2.0960
11	0.0857	1430.1590	2.0709
12	0.0899	1382.2637	2.0328
13	0.0923	1338.6662	**2.0275**
14	0.0923	1291.1924	2.1192

Табела 4: Резултати Хијерархијског кластеровања за различите вредности броја кластера, k

За визуализацију резултата приказан је **дендрограм** (*слика 9*), који илуструје хијерархијску структуру кластера и оптималан број кластера на основу силуете:



Слика 9: Дендрограм хијерархијског кластеровања

На основу различитих метрика, добијени су следећи закључци:

- Ниске вредности Силуета коефицијента и високе Дејвис-Болдин вредности указују да подаци немају јасну кластерску структуру.
- Различите метрике указују на различите оптималне вредности k (**3** за **Силуету**, **2** за **Калински-Харабаш, 13** за **Дејвис-Болдин**), што додатно показује недостатак јасних кластера.

На основу ових резултата, закључујемо да хијерархијско кластеровање није погодан алгоритам за ову врсту кластеровања.

3.1.4. GMM

Гаусов мешовит модел (енг. Gaussian Mixture Model - GMM) кластеровања примењен је на β-метилацијске вредности CpG локуса. GMM је пробабилистички модел који претпоставља да су подаци генерисани мешавином неколико нормалних расподела са различитим срединама и коваријансама.

Осим стандардних метрика које користимо да оценимо кластеровање: силуете, Калински-Харабаш индекса и Дејвис-Болдин индекса, за GMM кластеровање додатно користимо:

- **Log-Likelihood**: мери колико добро модел објашњава податке. За GMM, log-вероватноћа изражава колико су вероватни подаци с обзиром на параметре модела (већа вредност је боља). Математички израз за **Log-Likelihood:**

$$log\ L(\theta) = \sum_{i=1}^{N} log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)$$

где је,

- x_i- и-ти податак,
- π_k - тежина кластера k,
- $\mathcal{N}(x_i | \mu_k, \Sigma_k)$ - густина вероватноће нормалне расподеле за кластер k,
- K - број кластера,
- θ - параметри модела.

Проблем: **Log-Likelihood** увек расте када додамо више кластера, што може довести до преприлагођавања.

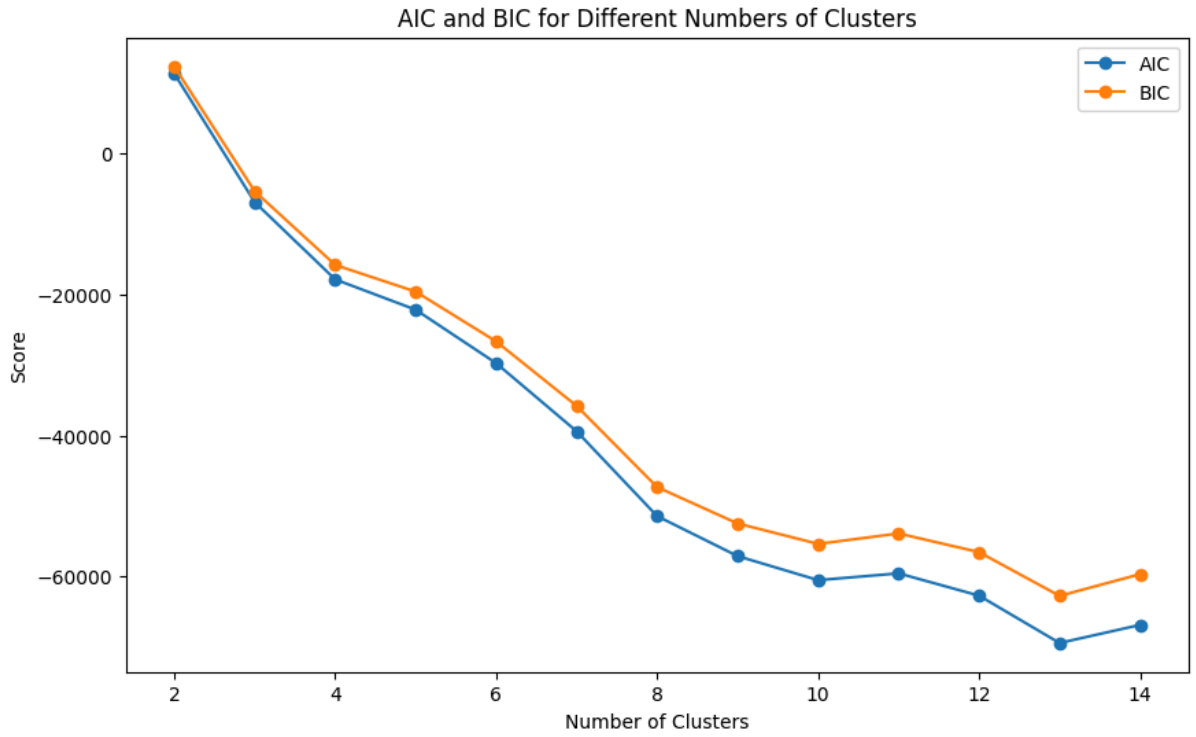
- **AIC (Akaike Information Criterion)**: балансира прецизност модела (log-вероватноћу) и његову сложеност (број параметара). Кажњава сложеније моделе. Мањи **AIC** значи бољи модел. Математички израз:

$$AIC = -2logL + 2K$$

- **BIC (Bayesian Information Criterion)**: Сличан AIC-у, али строже кажњава сложеније моделе, узимајући у обзир величину података N. Фаворизује једноставније моделе. Ако је N велико, пенализација сложености је јача, што спречава преприлагођавање. Математички израз:

$$BIC = -2logL + KlogN$$

Визуелно су приказани **AIC и BIC резултати** у зависности од броја компоненти, како би се видело где се јавља минимум ових критеријума (*слика 10*).



Слика 10: AIC и BIC вредности за GMM кластеровање

На основу различитих метрика, добијени су следећи закључци:

- **2 кластера**: Метрике као што су **силуета (0.1238), Калински-Харабаш индекс (2913.4824) и Дејвис-Болдин индекс (2.3187)** сугеришу да би подаци могли бити подељени у две групе, али кластери нису јасно раздвојени.
- **13 кластера**: **AIC (-69459.1287), BIC (-62770.4394) и Log-Likelihood (35586.5643)** указују да подаци могу подржавати детаљнију поделу на 13 кластера, али овај број може резултирати преприлагођавањем или подацима који су више вођени шумом него природном кластерском структуром.
- Ниједна метрика не указује на изузетно јаку кластерску структуру, што може значити да подати нису природно кластеровани.

На основу ових анализа, можемо закључити да GMM модел није оптималан за овај задатак кластеровања.

3.1.5. Спектрално кластеровање

Спектрално кластеровање је техника која користи спектралну анализу графова за груписање података.
Спектралном анализом бета вредности узорака, добијено је кластеровање чији је квалитет приказан у табели 7. Израчунати коефицијенти упућују на то да кластери нису јасно раздвојени. Иако је Дејвис - Булдин коефицијент релативно добар за предложених 9 кластера, ниске вредности силуета коефицијента и Калински - Харабаш индекса указују да спектрално кластеровање није најбољи алгоритам кластеровања за коришћене податке.

3.2. Клизни прозор

Клизни прозори представљају технику у којој се CpG региони анализирају **у континуираним сегментима фиксне дужине** дуж хромозома. Ово омогућава локално груписање CpG региона, уместо анализе појединачних CpG места.

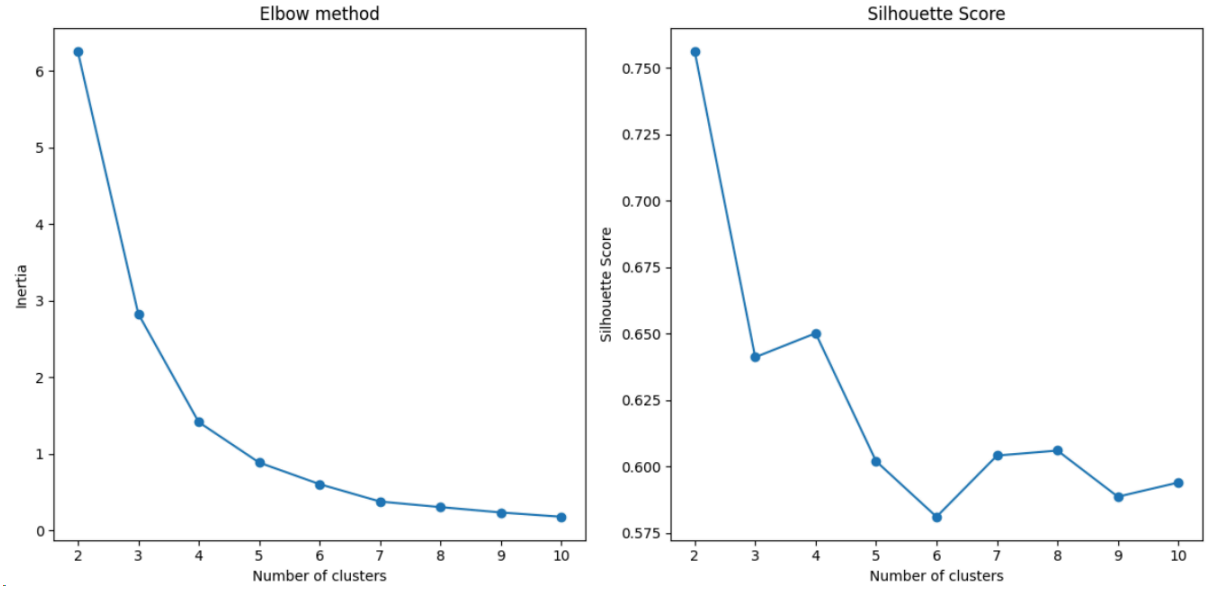
Податке смо поделили у прозоре величине 16,330 bp (парова база). Иако ова величина није резултат експлицитне оптимизације, она представља компромис између биолошке релевантности и статистичке стабилности: са једне стране, прозори овакве величине су довољно мали да задрже локалну специфичност CpG обележја, односно да обухвате CpG локусе који потенцијално учествују у истом регулаторном механизму, а са друге стране, прозори су довољно велики да садрже довољан број CpG локуса, што омогућава рачунање унутарпрозорске корелације на статистички валидан начин.

Сваки CpG регион је додељен одговарајућем прозору на основу своје геномске координате.
Унутар сваког прозора, анализирали смо **просечну корелацију CpG локуса** како бисмо детектовали регије са значајним разликама у нивоима метилације.

Један од циљева ове анализе је проналажење CpG региона са наглим променама у обрасцима метилације. Проналазимо CpG регионе у којима се метилација мења у односу на суседне регионе.
Овај приступ омогућава **идентификацију ширих епигенетских образаца.**

3.2.1. KMeans

Употребом KMeans алгоритма извршено је кластеровање података (описаних у уводном делу овог поглавља) који представљају корелацију метилационих профила (односно бета вредности) између узорака у једном клизном прозору. Резултати овог алгоритма приказани су у табели 8. Поред до сада виђених метрика, применили смо и лакат методу (енг. *elbow*) како бисмо одредили оптималан број кластера. На наредној слици приказани су, поређења ради, графици добијеном примене ове методе, као и вредности силуета коефицијента за различит број кластера (*слика 11*).



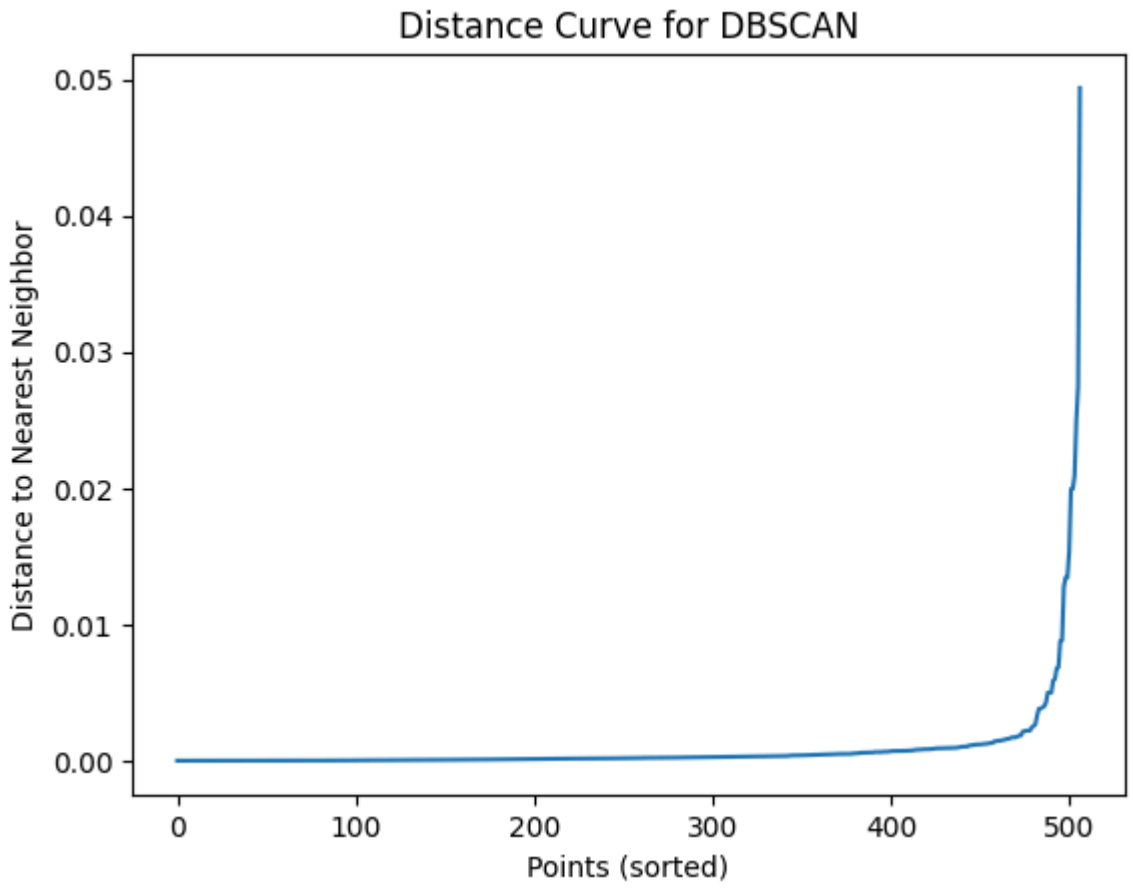
Слика 11: Графици вредности метрика добијених KMeans кластеровањем

Као што је већ познато, оптималан број кластера који сугерише лакат метода представља ону тачку на графику након које је смањење инерције незнатно. У нашем случају, за тај број бисмо могли да узмемо 7. То се значајно разликује од вредности коју препоручује силуета коефицијент и која износи 2. Оваква разликовања показују да кластеровање наших података није најбоље реализовати употребом овог алгоритма.

3.2.2. DBSCAN

У овој анализи, DBSCAN је примењен на регионе CpG локуса добијене клизним прозором, а оптимални параметри су одређени следећим поступком:

- Одређивање оптималне вредности **eps** (радијуса густине):
  - Израчунава се удаљеност до најближег суседа за сваку тачку.
  - Користи се **крива удаљености (Distance Curve)** (*слика 12*) да би се визуелно детектовала оптимална вредност `eps`.
  - *KneeLocator* се користи за аутоматско одређивање "лакат" тачке.



Слика 12: Крива удаљености

График приказује удаљеност до најближег суседа за сваку тачку у растућем редоследу, како бисмо идентификовали *eps* где се дешава нагли пораст удаљености. Добијена оптимална вредност *eps* је 0.0199.

- Одређивање оптималне вредности **`min_samples:`**
  - Примењен је DBSCAN са фиксном вредношћу **eps = 0.0199**, док се параметар `min_samples` тестира у опсегу [3, 14].

Добијени резултати кластеровања за различите вредности min_samples :

`min_samples`	**Силуета коефицијент**	**Калински-Харабаш индекс**	**Дејвис-Болдин индекс**	**k**
3	0.6037	307.2693	0.7997	7
4	0.6037	307.2693	0.7997	7
5	0.5989	363.7460	0.5921	6
6	0.5760	368.5578	0.5567	7
7	0.5378	347.3726	1.2247	7

8	0.5778	479.5874	2.6394	5
9	0.6723	624.3237	1.4568	4
10	0.6723	624.3237	1.4568	4
11	**0.7172**	933.0871	**0.3621**	3
12	0.7172	933.0871	0.3621	3
13	0.4891	752.7100	0.4537	4
14	0.7082	**1024.2194**	0.3962	3

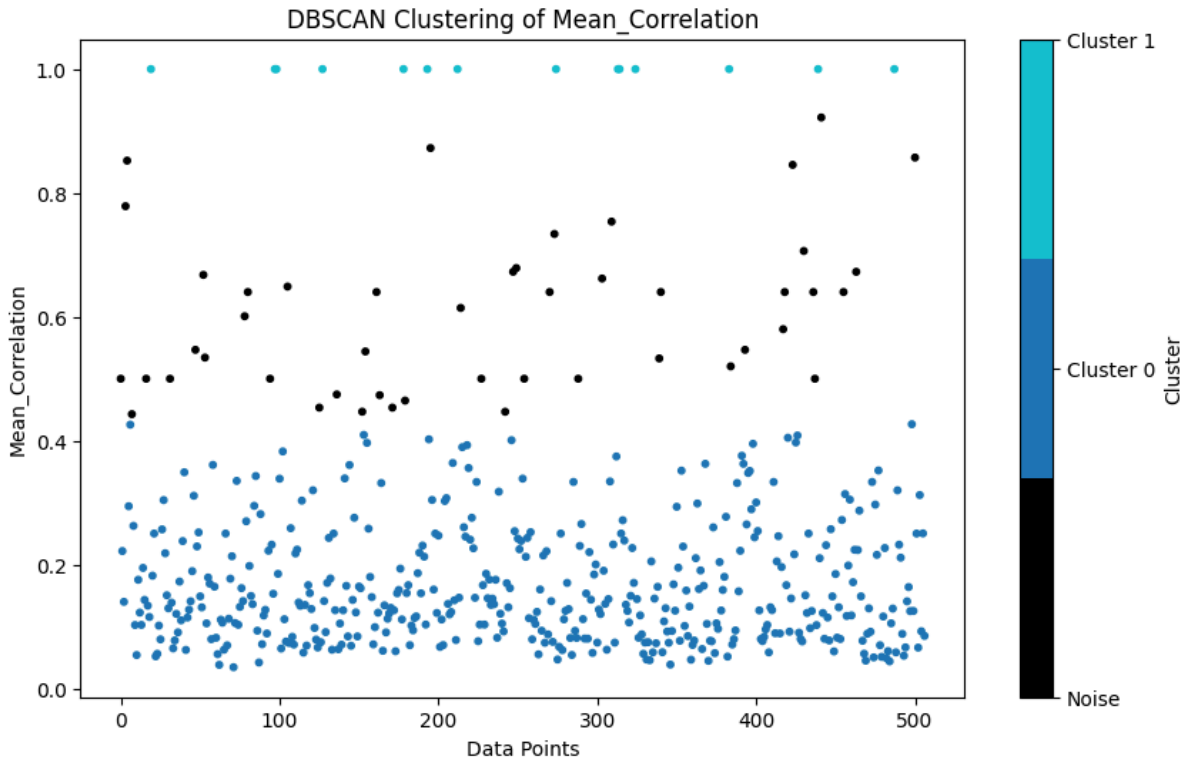
Табела 5: Резултати DBSCAN кластеровања за различите вредности min_samples

На основу различитих метрика, добијени су следећи закључци:

- **Најбољи Силуета коефицијент**: min_samples = 11 (0.7172), **3 кластера**
- **Најбољи Калински-Харабаш индекс**: min_samples = 14 (1024.2194), **3 кластера**
- **Најнижи Дејвис-Болдин индекс**: min_samples = 11 (0.3621), **3 кластера**

За све три метрике, оптималан број кластера је 3, што снажно сугерише да подаци природно формирају три различите групе.
Резултати истичу важност min_samples параметра, пошто и 11 и 14 дају најбоље резултате метрика метрике. Нешто нижа вредност од 11 обезбеђује најбоље укупно класетовање на основу силуете и Дејвис-Булдиновог резултата.

На основу резултата примењујемо DBSCAN са **eps = 0.0199** и **min_samples = 11.** Затим визуелизујемо кластеровање тако што приказујемо Mean_Correlation вредност по кластерима (*слика 13*):



Слика 13: Визуализација кластера клизним прозором

DBSCAN је идентификовао густо груписане регионе и одвојио тачке које нису део јасних кластера као "шум" (-1).

3.2.3. Хијерархијско кластеровање

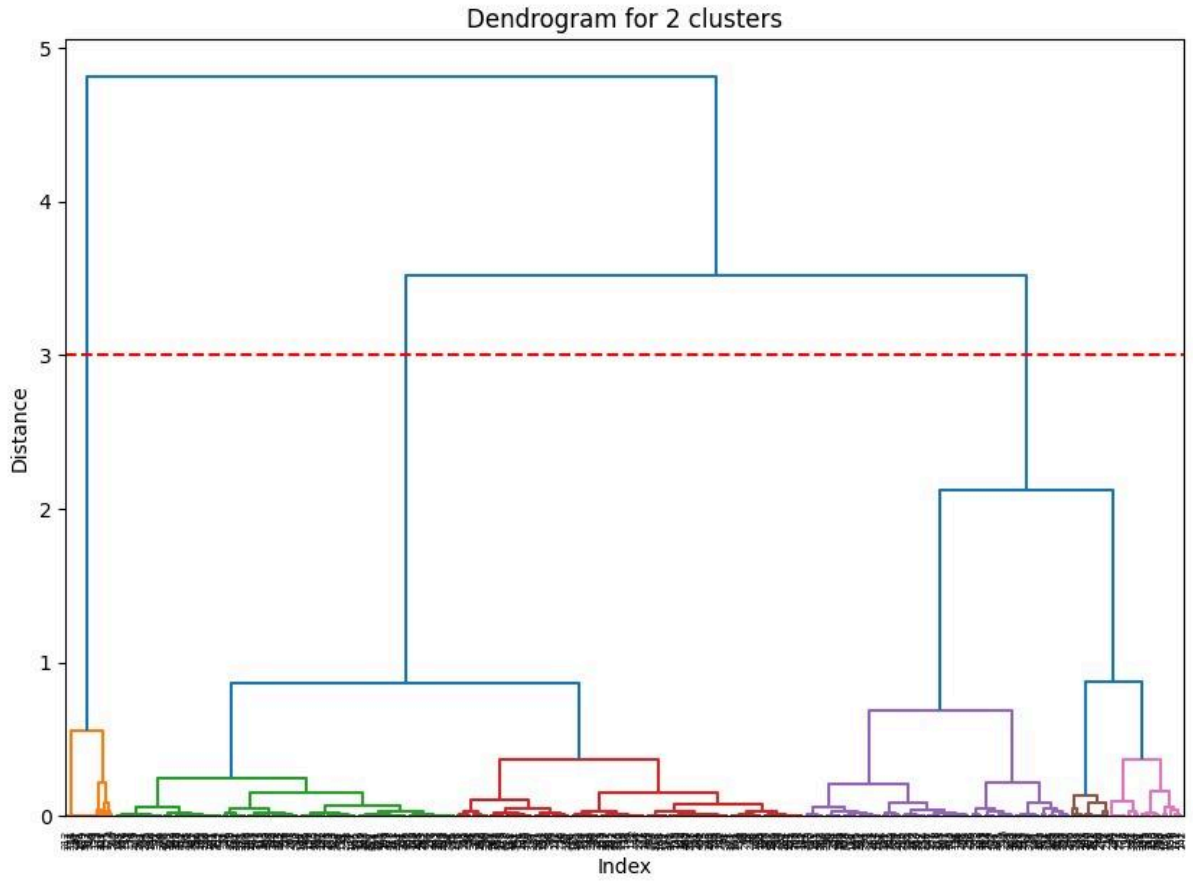
Извршено је хијерархијско кластеровање података употребом Вардове методе (енг. *Ward*), која минимизује укупну варијансу унутар кластера. Разматран је различит број кластера (од 2 до 10), и за сваки од њих израчунате су вредности три различите метрике, које су приказане у табели 6.

k	Силуета коефицијент	Калински-Харабаш индекс	Дејвис-Болдин индекс
2	**0.7690**	589.5474	**0.2538**
3	0.6124	1202.8096	0.5037
4	0.6374	2279.0606	0.4744
5	0.6351	2345.5475	0.4279
6	0.5663	2885.8821	0.4501
7	0.5868	3589.9921	0.4558
8	0.5934	4499.9564	0.4523

9	0.5976	4942.8308	0.4538
10	0.5819	**5875.6253**	0.4503

Табела 6: Резултати хијерархијског кластеровања за различите вредности броја кластера, k

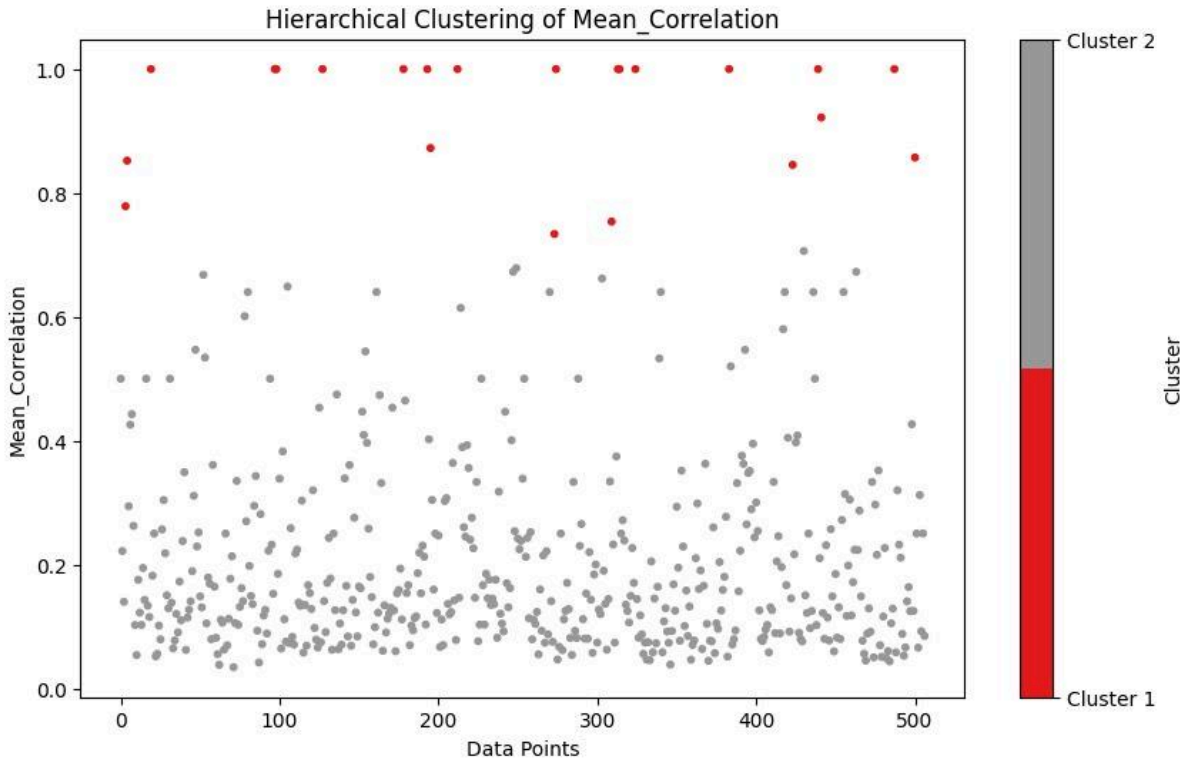
За визуализацију резултата приказан је **дендрограм** (*слика 14*), који илуструје хијерархијску структуру кластера и оптималан број кластера на основу силуете:



Слика 14: Дендрограм хијерархијског кластеровања

На основу израчунатих метрика, закључујемо да висока вредност силуета коефицијента показује да се релативно јасно могу уочити две кластерске структуре (ту чињеницу додатно подржава и Дејвис - Болдинов индекс). Иако Калински - Харабашов индекс препоручује 10 кластера, високе вредности силуета коефицијента и ниске вредности Дејвис - Болдиновог указују да је оптималан број кластера 2 и да су ови кластери добро раздвојени.

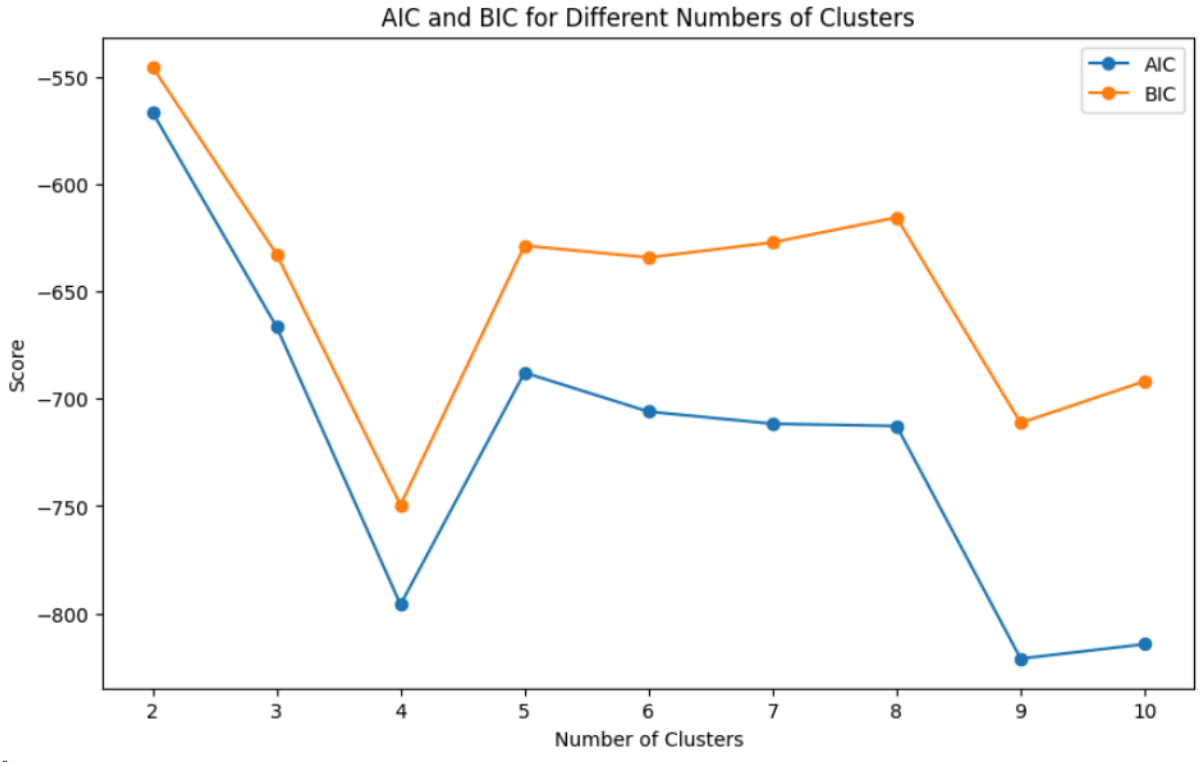
Кластеровање визуелно представљамо тако што приказујемо расподелу вредности *Mean_Correlation* за сваки кластер (*слика 15*).



Слика 15: Визуализација кластера клизним прозором

3.2.4. GMM

Начин функционисања GMM алгоритма, као и додатне метрике квалитета добијених кластера, објашњени су у одељку 3.1.4. На исти начин израчунате су вредности метрика и у овом случају, а на *слици 16* приказани су AIC и BIC резултати у зависности од броја компоненти, како би се видело где се јавља минимум ових критеријума (односно који је оптималан број кластера).

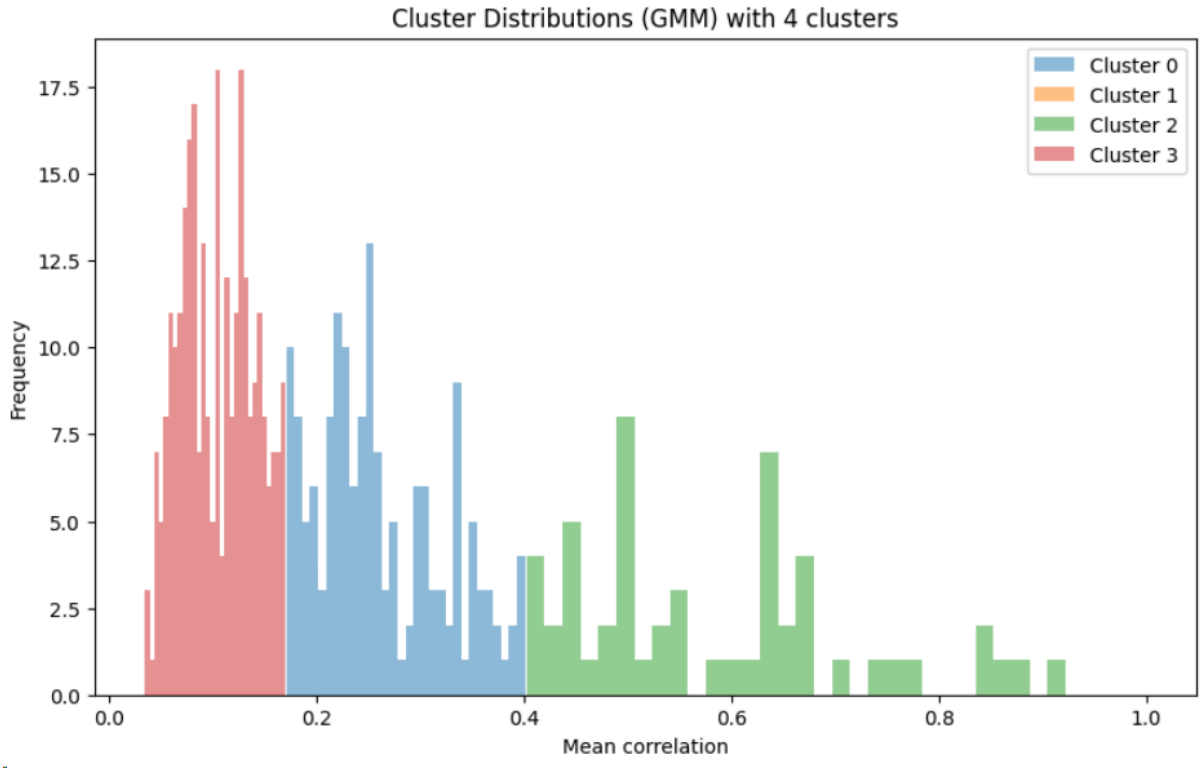


Слика 16: AIC и BIC вредности за GMM кластеровање

Иако силуета коефицијент, Дејвис - Болдин и Калински - Харабаш индекси сугеришу да је оптималан број кластера 2, најмање вредности AIC и BIC постижу се за 4 кластера. То може да значи да у скупу података постоје две веће подгрупе података, и да (пошто AIC и BIC вредности указују на 4 кластера) важи нешто од следећег:

- У оквиру ових подгрупа постоје релативно добро раздвојене мање групе
- Преприлагођавање уколико додатни кластери непотребно раздвајају две веће подгрупе

За четири кластера која су предложили AIC и BIC, испитана је расподела података у оквиру кластера. Хистограми тих расподела приказани су на *слици 17*:



Слика 17: Хистограм расподеле података по кластерима

3.2.5. Спектрално кластеровање

У овој анализи спектрално кластеровање примењено је на регионе CpG локуса добијене клизним прозором.
Користећи параметарску претрагу, *ParametarGrid* из библиотеке ***scikit-learn***, испитујемо број кластера, *k,* у опсегу [2, 14], како бисмо пронашли оптималан број кластера у подацима.

На основу различитих метрика, добијени су следећи закључци:

- Негативан **Силуета Коефицијент (-0.0161)** и низак **Калински-Харабаш индекс (1.3719)** указују да спектрално кластеровање не успева да пронађе значајне кластере у овом скупу података.
- **Дејвис-Болдин индекс (0.6350)** даје најнижу вредност за k = 13, али је недоследан у односу на друге метрике.
- Квалитет кластеровања зависи од конструкције графа сличности, а у овом случају, изгледа да "nearest_neighbors" афинитет није добро дефинисан.

На основу ових анализа, можемо закључити да спектрално кластеровање није одговарајући алгоритам за овај задатак кластеровања.

4. Дискусија

На основу метрика кластеровања по CpG локусима приказаних у Табели 7, резултати указују на следеће:
KMeans алгоритам је имао најбоље резултате по Калински-Харабаш индексу и релативно најмањи Дејвис - Булдин индекс, али је силуета коефицијент био веома низак (али и даље већи од свих осталих примењених алгоритама).

Ово сугерише да CpG локуси посматрани појединачно немају изражену природну кластерску структуру, што је у складу са визуелном анализом помоћу PCA (*поглавље 2*).

	Коефицијенти / Оптималан број кластера		
Алгоритам	**Силуета**	**Калински - Харабаш**	**Дејвис - Булдин**
KMeans	0.1888 / 2	4175.5390 / 2	1.7309 / 14
DBSCAN	0.0332 / 3	588.3179 / 3	2.2992 / 3
Хијерархијско кластеровање	0.1357 / 3	2754.2241 / 2	2.0275 / 13
GMM	0.1238 / 2	2913.4824 / 2	2.3187 / 2
Спектрално кластеровање	-0.0161 / 2	1.3719 / 7	0.6350 / 9

Табела 7: Метрике квалитета кластеровања по CpG локусима

Ситуација се значајно мења у Табели 8, где су приказане вредности метрика кластеровања клизним прозором.

Хијерархијско кластеровање је постигло највише вредности по свим метрикама, указујући да је оптималан број кластера 2. Ово указује на јасно дефинисану кластерску структуру у регионима добијеним клизним прозором.

DBSCAN и KMeans су такође дали добре резултате, што указује на постојање два или три густинска кластера у регуларним CpG регионима.

Спектрално кластеровање, иако нешто боље него у случају кластеровања по CpG локусима, није дало јасну кластерску сегрегацију.

Ово показује да CpG региони добијени клизним прозором садрже много израженије кластерске обрасце, што оправдава примену овог приступа у епигенетским анализама.

	Коефицијенти / Оптималан број кластера		
Алгоритам	**Силуета**	**Калински - Харабаш**	**Дејвис - Булдин**
KMeans	0.7562 / 2	6540.8118 / 10	0.5135 / 3
DBSCAN	0.7172 / 3	1024.2194 / 3	0.3621 / 3
Хијерархијско кластеровање	0.7690 / 2	5875.6253 / 10	0.2538 / 2
GMM	0.5738 / 2	5750.7647 / 2	0.4278 / 2
Спектрално кластеровање	- 0.0161 / 13	1.3719 / 8	0.6350 / 13

Табела 8: Метрике квалитета кластеровања клизним прозором

5. Закључак

У овом раду спроведена је кластер анализа епигеномских података добијених са HumanMethylation450 чипа, са циљем да се истраже обрасци ДНК метилације код серозног цистаденокарцинома јајника. Анализа је обухватила два различита приступа кластеровању:

- кластеровање појединачних CpG локуса
- кластеровање региона добијених клизним прозором, односно обједињених геномских области које садрже више CpG места.

Метилација ДНК је један од главних механизама епигенетске регулације и подразумева везивање метил-група на CpG динуклеотиде. Вредности метилације у раду су изражене као β-вредности, које се крећу у интервалу од 0 до 1 и представљају степен метилације на одређеној CpG локацији.

Коришћени су различити алгоритми за кластеровање - KMeans, DBSCAN, хијерархијско кластеровање, GMM и спектрално кластеровање, а евалуација квалитета кластера обављена је на основу више метрика: силуета коефицијента, Калински-Харабаш индекса и Дејвис-Болдин индекс.
За GMM модел додатно су коришћене AIC, BIC и Log-Likelihood метрике, које се заснивају на статистичком моделовању и оцењују квалитет прилагођености нормалних расподела.

Резултати добијени у овом раду показују да је кластеровање CpG региона добијених клизним прозором значајно ефикасније у откривању структуре података у поређењу са кластеровањем појединачних CpG локуса.

- Од свих примењених алгоритама, хијерархијско кластеровање је дало најбоље резултате на клизним прозорима, што указује на постојање природних епигенетских групација унутар генома.
- Кластеровање појединачних CpG локуса, с друге стране, показало је слабе резултате у свим методама, што сугерише да се CpG обрасци боље анализирају у контексту регионалне структуре.

Ови налази подржавају примену регионалне (window-based) епигенетске анализе и могу допринети бољем разумевању епигенетских потписа код серозног цистаденокарцинома јајника. Даља истраживања могу укључити комбинацију метилацијских и експресионих података, као и примену супервизованих метода за класификацију туморских подтипова.

6. Референце

1. Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes & Development, 16(1), 6–21. https://doi.org/10.1101/gad.947102
2. Valente, A., Vieira, L., Silva, M. J., & Ventura, C. (2023). The Effect of Nanomaterials on DNA Methylation: A Review. *Nanomaterials*, *13*(12), 1880. https://doi.org/10.3390/nano13121880
3. Rasuli B, Niknejad M, Yap J, et al. Ovarian serous cystadenocarcinoma. Reference article, Radiopaedia.org (Accessed on 31 Jan 2025) https://doi.org/10.53347/rID-14495
4. Du, P. et al. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics, 11(1), 587. https://doi.org/10.1186/1471-2105-11-587
5. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
6. Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1–27. https://doi.org/10.1080/03610927408827101
7. Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909