Metode raspoznavanja obrazaca i njihova primjena u analitičkoj kemiji

Rakas, Anja

SVEUČILIŠTE U ZAGREBU
FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE
SVEUČILIŠNI PREDDIPLOMSKI STUDIJ

Anja Rakas

ZAVRŠNI RAD

Zagreb, rujan 2018.

SVEUČILIŠTE U ZAGREBU
FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE
SVEUČILIŠNI PREDDIPLOMSKI STUDIJ

Anja Rakas

METODE RASPOZNAVANJA OBRZACA I NJIHOVA PRIMJENA U ANALITIČKOJ KEMIJI

ZAVRŠNI RAD

Voditelj rada: izv. prof. dr. sc. Šime Ukić

Članovi ispitnog povjerenstva: izv. prof. dr. sc. Šime Ukić
dr. sc. Mirjana Novak Stankov
prof. dr. sc. Irena Škorić

Zagreb, rujan 2018.

Zahvaljujem se mentoru izv. prof. dr. sc. Šimi Ukiću na brojnim savjetima, strpljenju i stručnoj pomoći pri izradi ovog završnog rada.

Također, zahvaljujem svojim roditeljima i bratu na bezuvjetnoj ljubavi, podršci i razumijevanju koje su mi ukazali tijekom studiranja.

Ovaj rad je izrađen u sklopu projekta „Modeliranje okolišnih aspekata napredne obrade voda za razgradnju prioritetnih onečišćivala“ Hrvatske zaklade za znanost na Fakultetu kemijskog inženjerstva i tehnologije Sveučilišta u Zagrebu.

SAŽETAK

Rudarenje podataka je novi pristup koji omogućava izvlačenje korisnih informacija iz skupova podataka. Prepoznavanje obrazaca izučava se u mnogim područjima kao što su biologija, medicina, psihologija, informatika, statistika... Postoje različite metode za prepoznavanje obrazaca. Primjerice linearna regresija, strojno učenje, metoda najbližeg susjeda i druge. Jedan od oblika prepoznavanja obrazaca je klasifikacija koja pokušava koja pokušava podijeliti ulazne veličine u različite skupove (klase).

U ovom radu opisani su načini prepoznavanja obrazaca i povezane metode koje su od značajne koristi analitičkoj kemiji.

Ključne riječi: rudarenje podataka, prepoznavanje obrazaca, metode, analitička kemija

SUMMARY

Data mining is a new approach of revealing valuable information from different and mostly large data sets. Pattern recognition is studding in many areas such as biology, medicine, psychology, informatics, statistics… There are different methods to recognize patterns. For example, linear regression, machine learning, closest neighbor method, and more. One of them is classification that attempts to divide the input value into various classes.

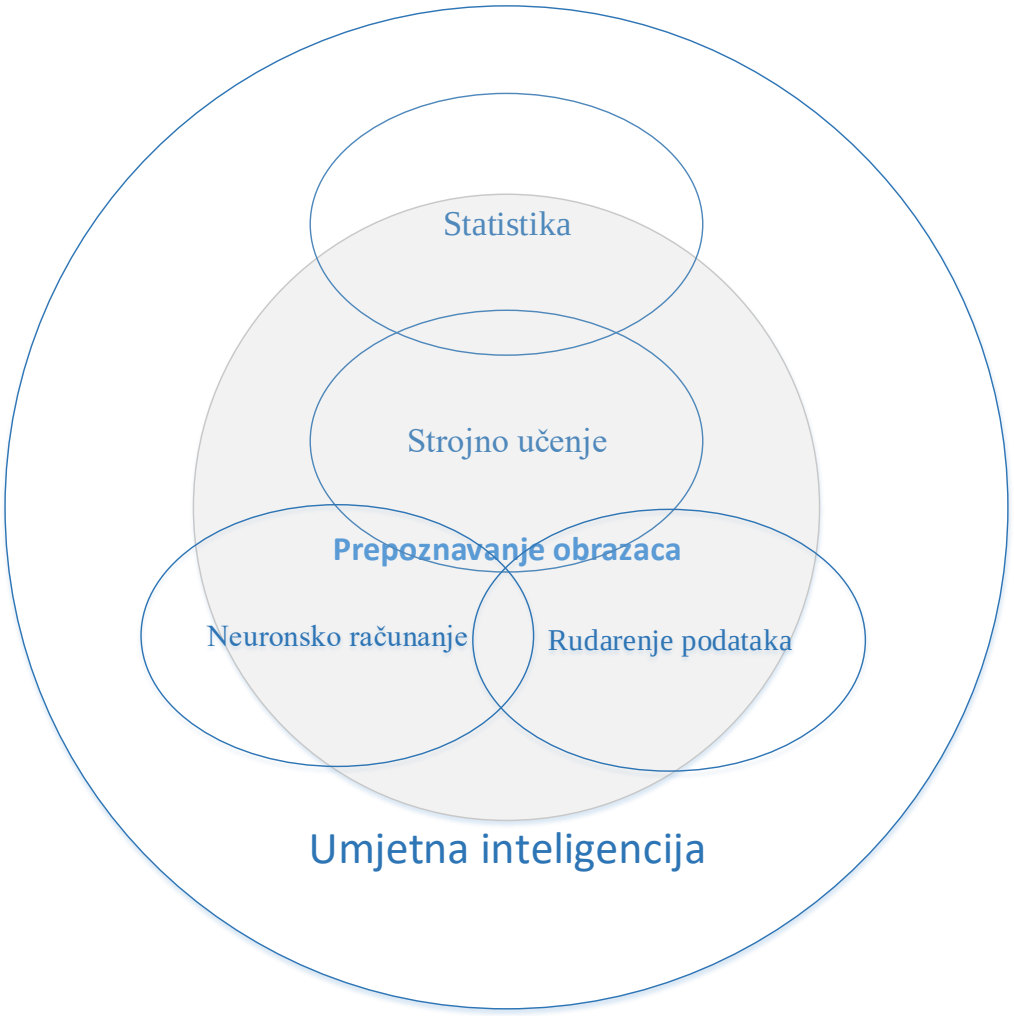
This work presents various modes of pattern recognition and related methods that are beneficial in analytical chemistry.

Key words: data mining, pattern recognitions, methods, analytical chemistry

SADRŽAJ

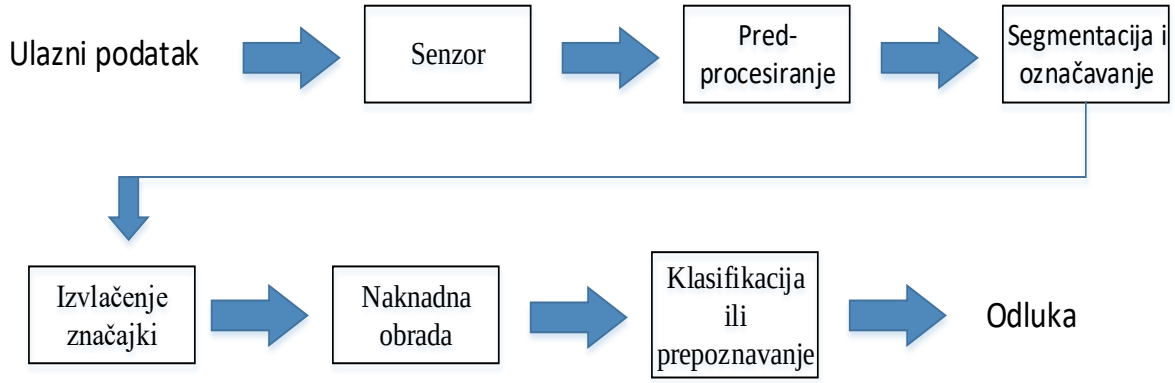
1. UVOD

Ljudi su toliko dobri u prepoznavanju objekata (ili uzoraka, obrazaca) da se ta sposobnost uzima zdravo za gotovo i teško je analizirati same korake u tom procesu. Općenito se čini da je lako razlikovati zvuk ljudskog glasa od zvuka violine, broj tri od broja osam, aromu nekog cvijeta i luka. Svakodnevno prepoznajemo lica koja vidimo oko sebe, no radimo to nesvjesno. Upravo zbog toga što ne možemo objasniti našu stručnost, teško nam je osmisliti računalni program koji može isto. Lice svake osobe je obrazac koji se sastoji od određene kombinacije struktura (oči, nos, usta, kosti, …) koje se nalaze na određenim mjestima na licu. Analizom uzoraka slika lica, program bi trebao biti u mogućnosti primijetiti obrazac specifičan za lice i identificirati (ili prepoznati) ga kao lice, kao člana određene kategorije ili klase koju već poznajemo. To se zove prepoznavanje obrazaca. Može postojati nekoliko kategorija (ili klasa) i mi moramo sortirati (klasificirati) određeno lice u određenu kategoriju. Tako se dolazi do pojma klasifikacije. Treba imati na umu da u terminu prepoznavanje obrazaca izraz obrazac podrazumijeva široko tumačenje. Uključuje sve predmete koje je potrebno klasificirati, npr. kruške, otiske prstiju, zvučne valove i druge. Kada se radi sa podacima u analitičkoj kemiji rijetka je zainteresiranost za neobrađenim signalima koji dolaze iz instrumenata već je naglasak na kemijski objašnjivim informacijama. Primjerice, područje kromatograma koji odgovara određenom spoju ili intenzitetu spektra poznate valne duljine navesti zašto je važno rudarenje podataka, da je to srodno područje sa prepoznavanje obrazaca i da prepoznavanje obrazaca nije moguće bez rudarenja podataka. Još jedan od pojmova koji su ključni u analitičkoj kemiji jest podatkovno rudarenje (engl. *data mining*) koje obuhvaća svaki način izvlačenja "zlata vrijednih" informacija iz baza podataka. To je proces istraživanja i analiziranja velike količine podataka s ciljem izdvajanja podataka na najbolji mogući način i preoblikovanje istih u razumljivu strukturu za daljnju uporabu. Na idućoj slici (slika 1.) vidljiv je položaj samog pojma prepoznavanja obrazaca i srodnih područja.



Slika 1. Prepoznavanje obrazaca i srodna kemometrijska područja[1]

Klasifikacija je često posljednji korak u općem procesu (slika 2.). Uključuje razvrstavanje objekata u zasebne klase. Uobičajeni sustav prepoznavanja obrazaca sadrži senzor, pred-procesni mehanizam (prije segmentacije), segmentacija i označavanje, mehanizam izvlačenja značajki, skup primjera (*training data*) koji su već klasificirani (*post-processing*) i algoritam klasifikacije. Korak ekstrakcije značajki smanjuje podatke mjerenjem određenih karakterističnih svojstava ili značajki (veličina, oblik, teksture) označenih objekata. Te značajke, ili vrijednosti, prosljeđuju se klasifikatoru koji ocjenjuje prikazane dokaze i donosi odluku u vezi s klasom kojoj svaki objekt treba biti dodijeljen, ovisno o tome jesu li vrijednosti tih značajki unutar ili izvan tolerancije te klase. Postupak se primjerice koristi u klasificiranju lezija kao benignih ili malignih. [1]



Slika 2. Općeniti sustav klasifikacije[1]

2. PODATKOVNO RUDARENJE

Tehnologije prikupljanja i skladištenja informacija omogućile su dostupnima ogromne količine podataka s primjenom u većini važnih sferi društvenog života, poput poslovnog svijeta, znanstvene i medicinske zajednice, javne uprave i drugih. Skup aktivnosti uključenih u analizu tih velikih baza podataka referira se na različite pojmove, kao što su rudarenje podataka, otkrivanje znanja, prepoznavanje obrazaca i umjetna inteligencija.

Rudarenje podataka definira se kao izvlačenje korisnih informacija iz velikog skupa podataka. Konkretno, pojam podatkovnog rudarenja ukazuje na proces istraživanja i analize baza podataka, obično znatne veličine, s ciljem izdvajanja relevantnog znanja i stjecanja značajnih pravila koja se mogu primijeniti u budućnosti. Rudarenje podataka dobiva sve veću ulogu u teorijskim i praktičnim studijama. Proces analize je po prirodi iterativan jer postoje različite faze koje bi mogle primijeniti povratne informacije i naknadne izmjene. Obično takav proces predstavlja suradnju između stručnjaka u praktičnoj primjeni i analitičara koji koriste matematičke modele za induktivno učenje. Doista, iskustvo pokazuje da podatkovno rudarenje zahtijeva česte intervencije analitičara kroz različite istraživačke faze i upravo zbog toga se ne može lako automatizirati. Također je nužno da je znanje stečeno ovakvim putem točno, odnosno mora biti potvrđeno podacima i ne voditi u pogrešno zaključivanje i donošenje pogrešnih odluka.

Pojam podatkovnog rudarenja se stoga odnosi na sveukupni proces koji se sastoji od prikupljanja i analize podataka, razvijanja induktivnih modela učenja i usvajanja praktičnih odluka i posljedičnih odluka temeljenih na stečenom znanju. Pojam matematičke teorije učenja rezerviran je za raznolikost matematičkih modela i metoda koje se mogu naći u jezgri svake analize podatkovnog rudarenja i koriste se za stvaranje novih znanja.

Proces obrade podataka temelji se na induktivnim metodama učenja, čija je glavna svrha izvući opća pravila počevši od skupa dostupnih primjera, koji se sastoje od prošlih zapažanja zabilježenih u jednoj ili više baza podataka. Drugom riječju, svrha *data mining* analize je doći do zaključaka počevši od obrazaca prošlih opažanja i generalizirati te zaključke s primjenom na cijelu populaciju. Zato je važno takvu analizu provesti na način da je točna koliko god je moguće. Modeli i uzorci identificirani na ovaj način poprimaju različite oblike, koji će biti opisani u daljnjem tekstu. Neki od tih su linearne jednadžbe, skupovi pravila u *if-then-else* obliku, klasteri, grafikoni i stabla.

Daljnja karakterizacija podatkovnog rudarenja ovisi o procedurama prikupljanja opažanja iz prošlosti i umetanja istih u baze podataka. Takvi zapisi obično se pohranjuju sa svrhom koja nije primarno potaknuta podatkovnim rudarenjem. Na primjer, informacije o kupnji od maloprodajne tvrtke, ili podaci o korištenju svakog telefonskog broja koje pohranjuju telefonske kompanije, u osnovi se bilježe u administrativne svrhe. Međutim, takvi podaci mogu se kasnije koristiti za korisnu analizu, na takav velik broj podataka može se primijeniti rudarenje podataka. Postupak prikupljanja podataka stoga je uglavnom neovisan od ciljeva podatkovnog rudarenja. Bitno se razlikuje od aktivnosti prikupljanja podataka koje su određene unaprijed definiranim shemama, što je karakteristično za klasičnu statistiku. U tom kontekstu, rudarenje podataka predstavlja sekundarni oblik analize podataka.

Aktivnosti rudarenja podataka mogu se podijeliti u dvije glavne istraživačke struje, prema glavnoj svrsi analize: interpretacija i predviđanje.

Interpretacija

Svrha tumačenja je identificirati redovne obrasce u podacima i izraziti ih kroz pravila i kriterije koje stručnjaci mogu lako razumjeti u vlastitoj domeni primjene. Generirana pravila moraju biti izvorna, originalna i ne-trivijalna kako bi zapravo povećala razinu znanja i razumijevanja određenog sustava interesa. Na primjer, za tvrtku u maloprodajnoj industriji moglo bi biti korisno grupirati one klijente koji su napravili kartice lojalnosti prema vlastitom kupovnom profilu. Tako generirani podaci mogu se pokazati korisnima u identificiranju novih tržišnih niša i usmjeravanju budućih marketinških kampanja te tvrtke.

Predviđanje

Svrha predviđanja je predvidjeti vrijednost koju će određena varijabla pretpostaviti u budućnosti ili procijeniti vjerojatnost budućih događaja. Na primjer, pružatelj telekomunikacijskih usluga može razviti analizu rudarenja podataka kako bi stekli prednost u odnosu na konkurenciju. Malo poduzeće može predvidjeti prodaju određenog proizvoda tijekom idućih nekoliko tjedana. Većina tehnika podatkovnog rudarenja temelje svoja predviđanja iz vrijednosti skupa varijabli koji su povezani s određenim entitetom u bazi podataka. Na primjer, model rudarenja podataka može ukazivati na vjerojatnost budućih pogodnosti za kupca, temeljenih na značajkama poput dobi, trajanju ugovora, postotku poziva

prema „drugim mrežama“ i slično. Postoje također i modeli temeljeni na vremenu, koji predviđaju putem prošlih vrijednosti varijabli područja interesa. [2,3]

3. METODE PREPOZNAVANJA OBRAZACA

U analitičkoj kemiji čest je problem odrediti ili predvidjeti svojstva objekata ili događaja koje nije moguće izravno mjeriti, ali se moraju izvesti iz neizravnih mjerenja. Primjeri su određivanje molekularnih struktura ili biološke aktivnosti nekog spoja. Ako je teorijski odnos između neizravnih mjerenja i nejasnih svojstava neadekvatno uspostavljen, tada metode prepoznavanja obrazaca mogu pružiti pristup rješavanju problema. Jedna od prvih i najpopularnijih priča o uspjehu kemometrije je prepoznavanje obrazaca. Mnogo kemije uključuje korištenje podataka za određivanje obrazaca. Metode prepoznavanja obrazaca primjenjuju se u svrhu klasificiranja nepoznatih objekata u kategorije ili za razdvajanje objekata u kategorije. Ako se napravi nekoliko mjerenja iz nekog objekta (npr. kemijskog spoja), dobiveni skup mjerenja koji pripada istom objektu smatra se obrascem ili "vektorom uzorka" u matematičkom obliku. Broj mjerenja definira broj dimenzija i bitan je parametar za metode prepoznavanja obrazaca.[1]

3.1. BINARNI KLASIFIKATORI

Statističari su razvili brojne matematičke metode prepoznavanja obrazaca koje se koriste u raznim područjima znanosti, medicine i tehnologije. Većina uspješnih metoda prepoznavanja obrazaca koje su predložene za kemijske probleme konceptualno su jednostavne; razumijevanje i primjena tih metoda ne zahtijevaju opsežno poznavanje matematike ili statistike.

Prepoznavanje obrazaca pomoću metode binarnih klasifikatora (engl. *binary classifiers*) koristi se da se vidi:

1. Postoji li dovoljno dokaza može li se dobiveni analitički podatak koristiti za određivanje je li obrazac član jedne od dvije predefinirane grupe.
2. Ako je, kojoj grupi obrazac pripada.[4]

Uzmimo jednostavan primjer gdje svaki objekt ima samo dva mjerenja x_1 i x_2. Svaki objekt se tada može prikazati točkom u dvodimenzionalnom koordinativnom sustavu. Ekvivalentni prikaz je vektor (vektor obrasca) od izvorišta do objekta. Pretpostavka za sve metode prepoznavanja obrazaca je da će se slični objekti pojavit blizu jedan drugome u prostoru obrasca, iako njihova sličnost nije kemijski mjerljiva. Vrlo očigledan slučaj prikazan je na

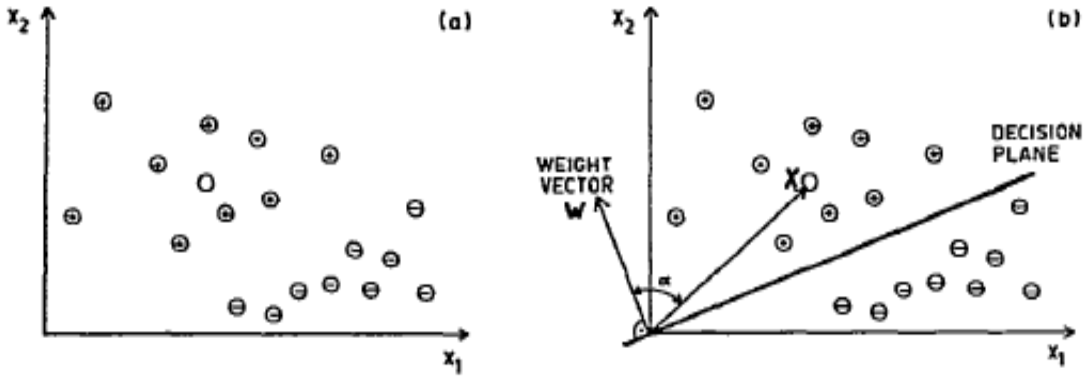
slici 3. Objekti formiraju dva različita klastera i svaki klaster sadrži samo objekte iste klase (istog razreda). Klasifikacija nepoznatog objekta (*o*) zahtijeva određivanje klastera kojem pripada ova točka. U stvarnim kemijskim primjenama nužan je prostor obrasca od puno više dimenzija, ne samo dvije ili tri. Klasteriranje u višedimenzionalnom prostoru, naravno, nije izravno vidljivo tako da je potrebno primijeniti posebne metode.

Sama spomen višedimenzionalnog prostora često je dovoljna za gašenje interesa za prepoznavanje obrasca, pa se treba naglasiti da ne postoji kvalitativna razlika između geometrije, recimo, stodimenzionalnog prostora i onog od dvije ili tri dimenzije. Razlika je samo kvantitativna, što ne predstavlja nikakav problem kada su dostupna računala.

Većina jednostavnih metoda klasifikacije može se objasniti dvodimenzionalnim primjerima. Proširenje na više dimenzija uglavnom je formalno i ne zahtijeva vizualizaciju više od tri dimenzije.

U binarnoj klasifikaciji moraju se razlikovati dvije isključive klase. Ako klase formiraju dobro razdvojene klastere, oni mogu biti potpuno odijeljeni "ravninom odlučivanja" (ravna linija u dvodimenzionalnom primjeru, *decision plane*); u ovom slučaju, skup podataka se linearno odvaja. Crta koja razdvaja klastere obično se definira vektorom odlučivanja (*weight vector*) koji je ortogonalan ravnini odlučivanja (slika 3.b). Taj vektor težine odlučuje pripada li točka klasi 1 ili 2. Kako bi se klasificirao vektor uzorka, potrebno je izračunati skalarni produkt (točku) *s* između vektora težine *w* i vektora obrasca *x*.

Slika 3.a prikazuje grupiranje objekata u prostoru obrasca gdje su x_1 i x_2 indirektna mjerenja i objekti su prikazani kao točke; klasifikacija nepoznatog objekta (*o*) zahtijeva određivanje klastera kojem pripada objekt (*o*). Slika 3.b prikazuje kojoj skupini pripada nepoznati objekt jer se nalazi između vektora težine (*weight vector*) i ravnine odlučivanja (*decision plane*). [5]



Slika 3. Objekti koji formiraju dva različita klastera. 3.a) grupiranje objekata u prostoru obrasca u kojem su objekti predstavljeni točkama i b) pripadnost nepoznatog objekta jednoj od skupina[5]

$$s = w \cdot x = |w| \cdot |x| \cdot \cos\alpha \qquad (1)$$

Predznak skalarnog produkta je pozitivan za klasu 1, jer je kut između dva vektora manji od 90 stupnjeva. Kosinus je također pozitivan. Skalarni produkt je negativan za klasu 2 jer je kosinus negativan. Jednostavan trik omogućava da ravnina odlučivanja uvijek prolazi kroz izvorište: svi vektori obrasca se povećavaju dodatnom komponentom (x_3 u dvodimenzionalnom primjeru) s istom konstantnom vrijednošću u svim obrascima. Računanje skalarnog produkta vektora koji imaju više od dvije dimenzije se lakše obavlja drugom formulom:

$$s = wx = w_1x_1 + w_2x_2 + \cdots w_dx_d \qquad (2)$$

gdje je (s) skalarni produkt (s < 0 za klasu 1; s > 0 za klasu 2), (w_i) komponente vektora težine, (x_i) komponente vektora obrasca i (d) broj dimenzija (uključujući dodatnu komponentu). Dakle, klasifikacija nepoznatog objekta zahtijeva samo izvršenje množenja i dodavanja produkata. Džepni kalkulator prikladan je za ove izračune. [5]

3.2. LINEARNA REGRESIJA

Modeli linearne regresije predstavljaju najpoznatiju obitelj regresijskih modela i temelje se na skupu hipoteza koje se sastoje od linearnih funkcija. Kao posljedica toga, funkcionalni odnos ($Y = f(X_1, X_2, \ldots, Xn)$) svodi se na:

$$Y = w_1X_1 + w_2X_2 + \cdots + w_nX_n + b = \sum_{j=1}^{n} w_jX_j + b \tag{3}$$

Ako postoji samo jedna nezavisna varijabla X = X_1 – to jest, n = 1 – model linearne regresije naziva se jednostavnim, pa imamo:

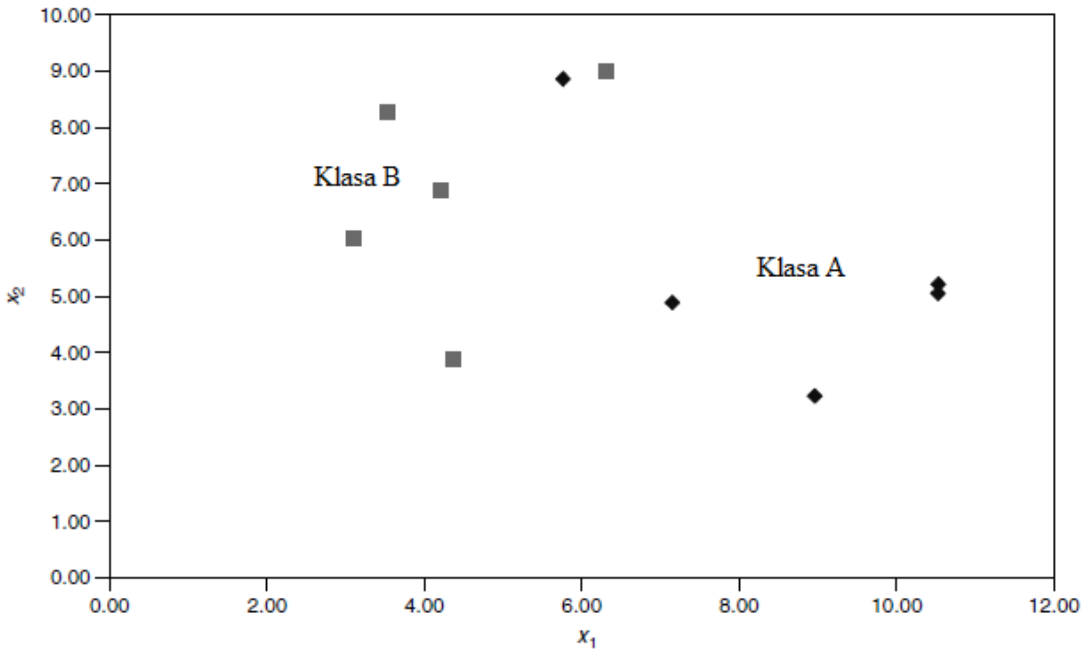
$$Y = wX + b \tag{4}$$

Geometrijsko tumačenje ovakvih modela čini ih posebno intuitivnim. Za jednostavne regresijske modele skup podataka smanjuje se na *m* parove vrijednosti (x_i , y_i) i $\in M$, koji su realizacije slučajnih varijabli *X* i *Y*. U prvoj fazi analize parovi vrijednosti mogu se grafički prikazati raspršenim nacrtom, kako bi se razvila vizualna percepcija bilo kojeg mogućeg odnosa koji postoji između slučajnih varijabli *X* i *Y*. [5]

3.3. METODA NAJBLIŽEG SUSJEDA

Metode SIMCA (engl. *Soft Independent Modeling of Class Analogy* , neovisno modeliranje analogije grupa), diskriminantne analize i DPLS (engl. *Discriminant partial least squers*) uključuju proizvodnju statističkih modela, kao što su glavne komponente i kanonske varijacije. Metode najbližih susjeda (engl. *K-nearest neighbor*, KNN) su konceptualno puno jednostavnije i ne zahtijevaju razrađene statističke proračune. KNN metodu kemičari koriste već više od 30 godina. Algoritam počinje s brojem objekata koji se dodjeljuju svakoj skupini, klasi. Na slici 4. se može vidjeti pet objekata koji pripadaju klasama A i B. Zabilježeni su pomoću dva mjerenja, koja mogu, primjerice, biti površine ispod kromatografskih pikova

(*chromatographic peak areas*) ili apsorpcijski intenziteti na dvije valne duljine. Grubi podaci prikazani su u idućoj tablici. (tablica 1) [6]



Slika 4. Objekti koji pripadaju klasama A i B [6]

Tablica 1. Primjer KNN klasifikacije: tri najbliže udaljenosti su podebljane [6]

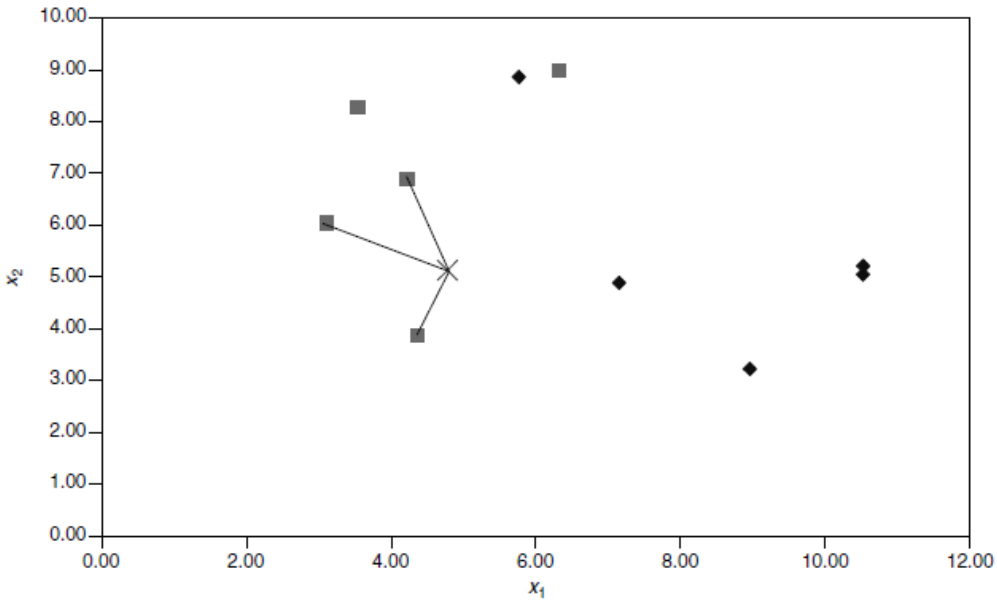
Klasa	**x_1**	**x_2**	**Udaljenost do nepoznatog objekta**	**Položaj**
A	5,77	8,86	3,86	6
A	10,54	5,21	5,76	10
A	7,16	4,89	2,39	4
A	10,53	5,05	5,75	9
A	8,96	3,23	4,60	8
B	3,11	6,04	**1,91**	**3**
B	4,22	6,89	**1,84**	**2**
B	6,33	8,99	4,16	7
B	4,36	3,88	**1,32**	**1**
B	3,54	8,28	3,39	5
Nepoznati objekt	4,78	5,13		

3.3.1. Metodologija

KNN-metoda se primjenjuje na slijedeći način:

1. Dodijeliti set za osposobljavanje poznatim klasama
2. Izračunati udaljenost nepoznatog objekta od svakog člana seta za osposobljavanje
   (pogledati tablicu 1). Obično se izračunava jednostavna Euklidova udaljenost.
3. Rangirati iste po redu (1 = najmanja udaljenost, i tako dalje)
4. Odabrati najmanje udaljenosti i pogledati kojim klasama (skupinama) je nepoznati objekt najbliži; taj broj je najčešće mali neparan broj. Slučaj gdje je K = 3 je ilustriran u slici. Sva 3 objekta pripadaju klasi B.
5. Uzima se "glas većine" i to se koristi za klasifikaciju. Primjerice, ako je K = 5, jedan od 5 najbližih objekata u tom slučaju pripada skupini A.
6. Ponekad je korisno koristiti KNN analizu za veći broj vrijednosti K, primjerice 3, 5, 7 i vidjeti mijenja li se klasifikacija. Na taj način mogu se otkriti nepravilnosti.

Ako postoji mnogo više od dva mjerenja, kao što je uobičajeno u kemiji, nužno je proširiti koncept udaljenosti u višedimenzionalni prostor. Iako ne možemo vizualizirati više od tri dimenzije, računala mogu računati geometriju u neodređenom broju dimenzija, a ideja udaljenosti ostaje jednostavna za generalizaciju. U slučaju prikazanom na slici 5. nije nužno izvršiti složeni račun za klasificiranje nepoznatog objekta, ali kada se radi o velikom broju mjerenja, primjerice u spektroskopiji, često je teško odrediti klasu nepoznatog objekta jednostavnim grafičkim prikazima.



Slika 5. Tri najbliža susjeda nepoznatom objektu [6]

3.3.2.Ograničenja

Ovakav konceptualno jednostavan pristup prikladan je za mnoge situacije, ali je važno razumjeti njegova ograničenja.

Prvo je da bi brojevi u svakoj klasi trebali biti približno jednaki, inače će „glasovi“ naginjati prema klasi koja ima više predstavnika (objekata). Drugo je da za najjednostavnije primjene svaka varijabla ima jednaku važnost. U spektroskopiji možemo snimiti stotine valnih duljina, od kojih će neke biti dijagnostički ili na neki drugi način povezane, a neke neće. Način kako ovo riješiti je ili odabir varijabli ili koristiti drugu mjernu udaljenost, primjerice kao u analizi klastera. Mahalanobisova udaljenost je uobičajena alternativna mjera. Treći problem je u tome što dvosmisleni ili vanjski uzorci u setu za osposobljavanje mogu dovesti do velikih problema krajnjoj, rezultantnoj klasifikaciji. Četvrto, metode ne uzimaju u obzir širenje ili odstupanje u klasama. Na primjer, ako bismo pokušali utvrditi je li forenzički uzorak krivotvorina, vjerojatno će klase u krivotvorini imati puno veće razlike nego klase nekrivotvorenih uzoraka.

Međutim, KNN je vrlo jednostavan pristup koji se lako razumije i programira. Mnogi kemičari vole takve pristupe, dok statističari često preferiraju više razrađene metode

temeljene na modeliranju podataka. KNN radi vrlo malo pretpostavki, dok metode temeljene na modeliranju često inherentno stvaraju pretpostavke koje nisu uvijek eksperimentalno opravdane, pogotovo kada se koriste statističke provjere za osiguranje vjerojatnosti pripadnosti nekoj skupini (klasi). U praksi, dobra strategija je koristiti nekoliko različitih metoda za klasifikaciju i vidjeti jesu li dobiveni slični rezultati. Često razlike u izvedbi različitih pristupa nisu u potpunosti posljedica samog algoritma, već su razlike u skaliranju podataka, mjerama udaljenosti, odabiru varijabli, metodi potvrde i tako dalje. Neki zagovornici određenih pristupa ne objašnjavaju ih u detalje.[6]

3.4. STROJNO UČENJE

Strojno učenje je nova rastuća tehnologija za prikupljanje znanja iz određenih podataka, i takvu tehnologiju mnogi ljudi počinju ozbiljno shvaćati.

Svijet je preplavljen podacima. Količina podataka u svijetu, u ljudskim životima, nastavlja rasti i ne vidi joj se kraj. Sveprisutna osobna računala olakšavaju spremanje stvari koje bi se inače bacile. Prijenosni hard diskovi s mogućnošću pohrane ogromne količine podataka, usb stickovi i razni drugi alati olakšavaju odgađanje odluka o tome što učiniti sa svim tim podacima. Vrlo jednostavno se čuva sve u elektronskom obliku. Elektronika bilježi odluke, izbore kupovine u supermarketima, financijske navike, dolaske i odlaske. Zapravo što god se radi nalazi se negdje u nekoj bazi podataka. WWW (*World Wide Web*) opskrbljuje nebrojenom količinom informacija i svaki izbor koji se napravi je zabilježen. Postaje se svjedok rastućem jazu između generacija podataka i razumijevanja istih. Kako se količina podataka neumoljivo povećava, proporcionalno tome ljudsko razumijevanje tih podataka se smanjuje. U svim tim podacima skriveno se nalaze informacije, potencijalno korisne, koje rijetko dođu na vidjelo ili se iskoriste.

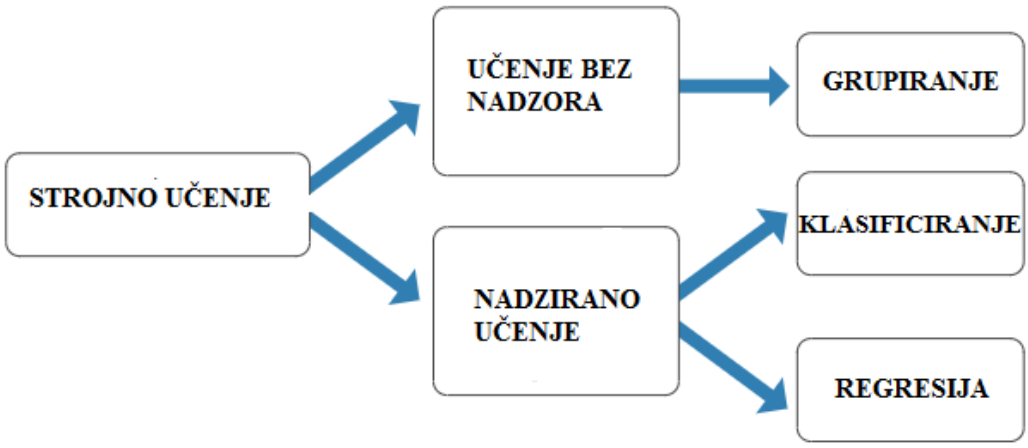
U podatkovnom rudarenju podaci se pohranjuju elektronički i pretraživanje je automatizirano, ili barem povećano, putem računala. Čak ni ovo nije osobito novo. Ono što je novo je zapanjujući porast u mogućnostima pronalaska obrazaca u podacima. Neobuzdan rast baza podataka u posljednjih nekoliko godina, i to baza podataka koje sadržavaju svakodnevne aktivnosti poput odabira kupaca u trgovinama, gura rudarenje podataka na čelo novih poslovnih tehnologija. Kako svijet raste u svojoj kompleksnosti, preplavljujući podacima koje proizvodi, rudarenje podataka postaje jedina nada za razjašnjavanje uzoraka koji se nalaze u

njemu. Analiziranje podataka na pravi, kvalitetan način je vrlo vrijedan resurs. Može dovesti do novih uvida i saznanja, otkrića, i u komercijalnom svijetu, nadasve bitne konkurentne prednosti.

Podatkovno rudarenje se bavi rješavanjem problema analizom podataka već prisutnih u bazama podataka. Definira se kao proces otkrivanja obrazaca u podacima. Proces mora biti automatski ili (češće) poluautomatski. Otkriveni uzorci moraju značajno voditi do određene prednosti, najčešće ekonomske.

Učenje poput inteligencije obuhvaća širok raspon procesa koji je teško precizno definirati. Mnoge tehnike u strujnom učenju (engl. *machine learning*) proizlaze iz nastojanja psihologa da na što precizniji način definiraju teorije o životinjskom i ljudskom učenju putem računalnih modela. Stroj uči kad god mijenja program ili podatke (na temelju ulaznih podataka ili kao odgovor na vanjske informacije), na način da se očekuje poboljšanje buduće učinkovitosti. Strojno učenje obično se odnosi na promjene u sustavima koji izvršavaju zadatke povezane umjetnom inteligencijom (engl. *artificial intelligence*).

Osnovne podjele modela strojnog učenja su modeli nadziranog strojnog učenja (*engl. supervised machine learning*) i modeli strojnog učenja bez nadzora (*engl. unsupervised machine learning*). Glavna razlika između navedenih modela je činjenica da nadzirani model zahtijeva znanje o vrijednosti podataka, odnosno tijekom učenja modela za predikciju model mora imati predanu i referentnu vrijednost koja proširuje informacije o tom podatku. Primjerice, kod predviđanje dizajna modela odjeće koji bi se ljudima mogli svidjeti na temelju dizajna koji su im se prethodno svidjeli, model prilikom učenja mora za svaki dizajn znati je li se on svidio korisniku ili nije. Modeli učenja bez nadzora najčešće se odnose na modele za grupiranje podataka temeljem sličnosti te detekciju anomalija u podacima. Primjer takvog modela može se naći u industriji gdje se na temelju podataka o sastavu nekog proizvoda mogu detektirati neispravnosti kako bi se isti uklonili iz uporabe. Na slici 6. prikazana je osnovna podjela modela strojnog učenja.



Slika 6. Tehnike strojnog učenja uključuju nadzirano učenje i učenje bez nadzora [7]

Sljedeća velika podjela modela strojnog učenja je podjela na klasifikacijske i regresijske modele. Svrha klasifikacijskih modela, kao što samo ime govori, jest dodijeliti određenu klasu nekom predmetu temeljem njegovih karakteristika. Primjer jednog takvog modela je predviđanje je li ispitanik sretan ili tužan temeljem njegovih objava na društvenim mrežama, gdje su klase najčešće predstavljene brojevima 0 i 1, ali imaju semantiku sretan i tužan. Regresijski modeli s druge strane rade s cijelim brojevima, gdje se primjerice predviđa prosječna plaća ljudi temeljem podataka o njihovom stupnju obrazovanja. Valja napomenuti da podjela modela s obzirom na vrstu predviđanja nije uzrokovana razlikama u samim modelima, već je ovisna o vrsti podataka i željenim rezultatima koje određeni model daje, to jest najčešće se isti model može jednako, ili s minimalnim promjenama, koristiti za obje vrste predviđanja.

Tijekom priprema sustava, razmatrano je nekoliko poznatih i više korištenih modela strojnog učenja. U ranoj fazi razmatrani su generalizirani linearni ili aditivni modeli, no s obzirom da takvi modeli zahtijevaju pretpostavku da se sve promatrane relacije ponašaju linearno u ovisnosti o faktorima, što je daleko od istine za složeni ekosustav, odustalo se od njihovog razmatranja. Zbog navedenih poteškoća linearnih modela bilo je potrebno potražiti nove naprednije modele. Prvi, često korišten model je regresijska analiza (engl. *Multiple Logistic Regression*). Iako spada u generalizirane linearne modele i dalje je često korišten i daje dobre rezultate zbog moći interpolacije kompliciranijih funkcija na linearne. Osim regresijske analize proučavani su i napredniji modeli strojnog učenja kao što su model najveće

regresije (engl. *maximum entropy model*, MAXENT), metoda potpornih vektora (engl. *support vector machines*), neuronske mreže (engl. *artificial neural network*), klasifikacijska i regresijska stabla (engl. *classification and regression trees*).

Inicijalni korak većine modela strojnog učenja jest normalizacija svojstava podataka. Naime s obzirom da svojstva mogu biti raznolika i raznih raspona vrijednosti potrebno ih je normalizirati kako svojstva najvećih vrijednosti ne bi previše utjecali na predikciju. Primjerice, u nekom hipotetskom primjeru moguće je kao svojstvo odabrati prosječnu plaću zaposlenika, što je vrijednost u rasponu nekoliko tisuća novčanih jedinica, i prosječan broj profesionalnih sportaša u poduzeću za koje se radi taj teorijski model. Prosječan broj profesionalnih sportaša vjerojatno je manji od 1, dok je plaća nekoliko desetaka puta veća vrijednost, stoga bi kod predikcije takvog modela ta vrijednost imala veći značaj i pretjerano utjecala na konačni rezultat. Kako bi se izbjegla situacija iz navedenog primjera vrijednosti se normaliziraju, to jest svode na raspon od 0 do 1, čime je smanjen utjecaj veće varijable na samu predikciju.

Ispravnost modela jedan je od najvažnijih faktora modela strojnog učenja. Dokazivanje ispravnosti modela vrši se testiranjem modela testnim podacima. U idealnom slučaju prilikom pribavljanja podataka koji će se koristiti za učenje modela potrebno je odvojiti određen broj podataka koji se mogu iskoristiti za testiranje. Testni podaci trebali bi pokriti dovoljan broj ekstremnih slučajeva kako bi što robusnije ispitali ispravnost modela. Željeni omjer testnih i trening podataka najčešće je 1:3, no u realnom slučaju to često nije moguće ostvariti. Najveći problem modela strojnog učenja upravo su podaci koji u stvarnosti ne mogu pokriti sve slučajeve i pripremiti model na sve kombinacije ulaznih parametara. Ipak, takva nesavršenost može se poprilično ispraviti naprednijim tehnikama interpolacije i uzorkovanja podataka, čime se greške modela dovode do zanemarivih postotaka ili do prihvatljivih postotaka s obzirom na broj podataka kojima se model učio. Broj podataka kojima se model uči trebao bi brojati milijune podataka, kako bi se što realnije pokrio svaki mogući slučaj te se takvi modeli mogu dovesti do preciznosti od čak 99%, iako takvo prikupljanje podataka zahtijeva dugi niz godina rada i provedbe statističkih analiza. U realnom slučaju gdje se model uči s desecima tisuća ili samo nekoliko tisuća podataka, dovoljno dobrom preciznošću smatra se sve iznad 85%, a nekada čak i manje.

S obzirom na probleme prilikom učenja i dokazivanja ispravnosti razrađena je dovoljno dobra metoda kojom se ovakvi modeli mogu testirati u realnom slučaju. Uzevši u obzir ograničenost broja podataka i potreban omjer, ideja znanstvenika je da se iz skupa svih

podataka kojim kreator modela raspolaže, uzme 25% za testne podatke, a 75% za trening podatke koji će se koristiti kod učenja modela. Na taj način se osigurava potreban omjer testnih i trening podataka. Nakon što se model nauči, testni podaci se provuku kroz metodu predviđanja te se dobiveni rezultat provlači kroz nekoliko mogućih validacijskih funkcija. S obzirom da testni podaci imaju prethodno poznatu željenu klasu (u slučaju klasifikacije) ili vrijednost (u slučaju regresije), validacija nije toliki problem. Najjednostavnija metoda validacije je računanje broja pogođenih predikcija te se taj broj podijeli s ukupnim brojem predikcija čime se određuje točnost modela. Ali, postoje i neke naprednije metode, od kojih je najpopularnije računanje greške metodom najmanjih kvadrata (engl. *root-mean-square-error*) čime je validacija robusnija na pogreške. Nakon što se dokaže ispravnost modela te dobije kvaliteta predikcije, točno se zna s kojom se razinom točnosti taj model može koristiti. Najčešće korištena metoda prepoznavanja obrazaca u kemiji je strojno učenje.[8]

3.5. ANALIZA KLASTERA

Analiza klastera dijeli podatke u skupine (klastere) koje su značajne i/ili korisne. Ako su nam cilj smislene skupine, tada bi klasteri trebali obuhvatiti prirodnu strukturu podataka. Međutim, u nekim slučajevima, analiza klastera je samo korisna polazna točka za druge metode, poput sažimanja podataka. Analiza klastera odigrala je važnu ulogu u širokom nizu područja, poput prepoznavanja obrazaca, pronalaženja informacija, strojnog učenja i rudarenja podataka.

3.5.1. Grupiranje za razumijevanje

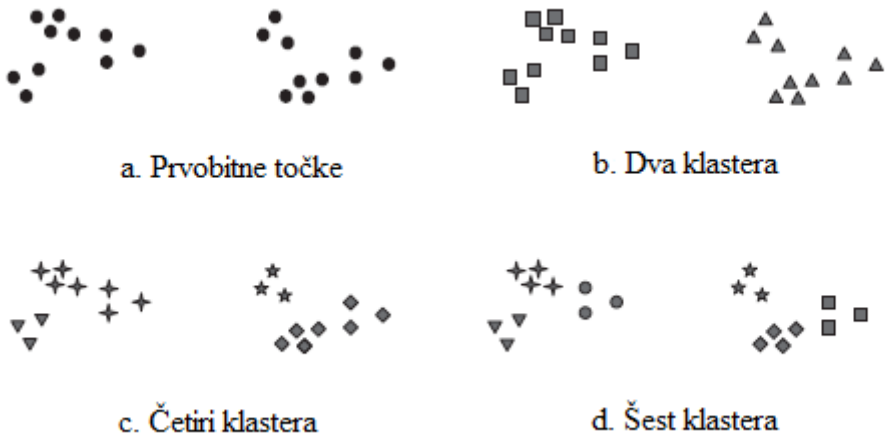
Klase, ili konceptualno smislene skupine objekata koje dijele zajedničke karakteristike, igraju važnu ulogu u tome kako ljudi analiziraju i opisuju svijet. Doista, ljudska bića su sposobna podijeliti objekte u skupine (grupirati) i dodjeljivati određene objekte tim skupinama (klasificirati). Na primjer, čak i relativno mala djeca mogu brzo prepoznati objekte na fotografiji poput zgrada, vozila, ljudi, životinja, biljaka i drugih. U kontekstu razumijevanja podataka, klasteri su potencijalne klase i analiza klastera je proučavanje tehnika za automatsko pronalaženje klasa.

3.5.2. Grupiranje za korisnost

Neke tehnike grupiranja karakteriziraju svaki klaster kao prototip, to jest podatkovni objekt koji predstavlja druge objekte u klasteru. Ti prototipovi klastera mogu se koristiti kao osnova za niz analiza podataka ili tehnika obrade podataka. Stoga je u kontekstu korisnosti analiza klastera zapravo proučavanje najreprezentativnijih prototipova klastera. Analiza klastera grupira podatkovne objekte temeljem informacija pronađenih u podacima koji opisuju objekte i njihove odnose. Cilj je da objekti unutar grupe budu slični (ili srodni) jedan drugome i različiti od (ili nepovezani) objekata u drugim skupinama. Što je veća sličnost (ili homogenost) unutar grupe i što je veća razlika između samih grupa, to je bolje ili jasnije grupiranje.

U mnogim primjenama, pojam klastera nije dobro definiran. Kako bi se bolje razumjele poteškoće prilikom odlučivanja što zapravo čini klaster, na slici 7. prikazano je dvadeset točaka i tri različita načina podjele tih točaka u klastere. Oblik točaka pokazuje

pripadnost nekom klasteru. Slike 7.a i b dijele podatke na dva, odnosno šest dijelova. Međutim, prividna podjela oba veća klastera u tri podskupine može jednostavno biti predmet ljudskog vizualnog sustava. Također, nije nerazumno reći da točke čine četiri klastera, kao što je prikazano na slici 7.c. Ova slika ilustrira da je definicija klastera neprecizna i da najbolja definicija ovisi o prirodi podataka i željenim rezultatima. [9]



Slika 7. Različiti načini grupiranja istog seta točaka [9]

Tehnike klasterskih analiza se bave istraživanjem skupova podataka kako bi se procijenilo mogu li se smisleno sažeti u smislu relativno malog broja grupa predmeta ili pojedinaca koji podsjećaju jedni na druge i koji su u nekim aspektima različiti od pojedinaca u drugim skupinama.

U većini primjena klasterske analize traži se razdvajanje podataka, u kojem svaki pojedinac ili objekt pripada jednom klasteru, a cijeli skup klastera sadrži sve pojedince. U nekim okolnostima, međutim, preklapajući klasteri mogu pružiti prihvatljivije rješenje. Također se mora zapamtiti da je jedan od odgovora koji se može dobiti analizom klastera taj da nije opravdano grupiranje podataka.

Osnovni podaci za većinu primjena klasterske analize su uobičajene $n \times p$ multivarijabilne podatkovne matrice, X, koje sadrže varijabilne vrijednosti koje opisuju svaki objekt koji treba biti grupiran:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix} \tag{5}$$

Varijable u matrici X često mogu biti mješavina kontinuiranog, rednog i/ili kategoričnog, a često će i nedostajati neki zapisi. Mješovite varijable i vrijednosti koje nedostaju mogu kompliciratigrupiranje podataka. U nekim primjenama, redovi matrice X mogu sadržavati ponovljena mjerenja iste varijable, ali primjerice u različitim uvjetima, različitim vremenima, ili na brojnim prostornim pozicijama itd. Jednostavan primjer u vremenskoj domeni pruža se mjerenjima, recimo, visine djece svakog mjeseca nekoliko godina. Takvi strukturirani podaci su od posebne naravi u tome što se sve varijable mjere u istom mjerilu, a klasterska analiza strukturiranih podataka može zahtijevati različite pristupe iz grupiranja nestrukturiranih podataka.

Analiza klastera u osnovi je otkrivanje skupina u podacima, a metode klasteriranja ne smiju se miješati s metodama dikriminacije i dodjele (u svijetu umjetne inteligencije koristi se pojam *supervised learning*), gdje su grupe unaprijed poznate i cilj analize je konstruirati pravila za klasificiranje novih pojedinaca u jednu ili u druge poznate skupine. [9]

4. PRIMJENA

Prepoznavanje obrazaca uglavnom se temelji na analiziranju skupova podataka. Koriste se kako bi se moglo odgovoriti na pitanja o temeljnom procesu ili eksperimentu. Prije korištenja računalnih metoda, važno je ispitati strukturu skupa podataka te odrediti koje su potrebne informacije. U nekim slučajevima informacije mogu biti istraživačke. Primjerice, može se istražiti postoji li dovoljno informacija u uzorku urina pacijenta kako bi se utvrdilo je li dijabetes napredovao. Ne zna se što je odgovor pa je cilj analize podataka prvenstveno istražiti podatke (treba utvrditi imaju li podaci predispozicije za ono što se traži). Zatim se testiraju hipoteze (koliko je sigurno da je isplativo analizirati te podatke). Čak i ako se zna da postoje bitne informacije (npr. može se razlikovati ljude prema spolu) ne zna se uvijek hoće li metoda analize (npr. ekstrakcija nakon koje slijedi kromatografija, masena spektrometrija ili NMR spektrometrija) biti adekvatna za otkrivanje relevantnih informacija. U drugim slučajevima želi se napraviti model predvidljivosti. Tada se može tvrditi (ili barem vjerovati) da su analitički podaci dovoljni za modeliranje informacija. Takav tip modeliranja naziva se predvidljivo modeliranje i uobičajeno je u tradicionalnoj analitičkoj kemiji; što je manja pogreška to je metoda bolja. Glavni cilj metode je smanjiti pogreške i povećati točnost. Obično, postoji značajan broj koraka potrebnih za pretvorbu sirovih podataka u informaciju koja je prikladna kao ulaz u algoritam prepoznavanja obrazaca, kao što su poboljšavanje razlučivosti, kalibriranje vrhova, koncentracije i tako dalje.

Podaci moraju biti oblikovani tako da postoji dovoljno informacija za proučavanje željenih čimbenika. Često postoji mnogo diskusija između statističara i kemometričara o važnosti eksperimentalnog plana. Mnogi skupovi podataka dostupni kemometričarima nisu savršeno obrađeni u formalnom statističkom smislu. To može značiti da se ne mogu upotrijebiti za izradu konačnih predviđanja. Bez obzira, iz tih podataka se mogu izvući bitne informacije. Zbog toga je važno za stručnjake kemometrije razumjeti ograničenja takvih skupova podataka. Ako se želi izbjeći pogreška potrebno je više konačnih rezultata ili veći skupovi podataka. Prije eksperimentalnog dijela to je nepoznato, a laboratorij može biti ograničen u smislu opremljenosti. Čak i ako se može vizualizirati takve eksperimente mogu se pojaviti novi problemi. Na primjer, uzorci će možda morati biti analizirani na kromatografu koji nije potpuno stabilan s vremenom, kolonu će trebati mijenjati i na taj način može doći do zastoja. Budući da se neki podaci ne mogu savršeno obraditi koristeći klasične statističke metode, mora postojati kompromis i kemometričar mora biti sistematičan. [4]

4.1. Slučaj 1: Analiza hrane bliskom infracrvenom spektroskopijom

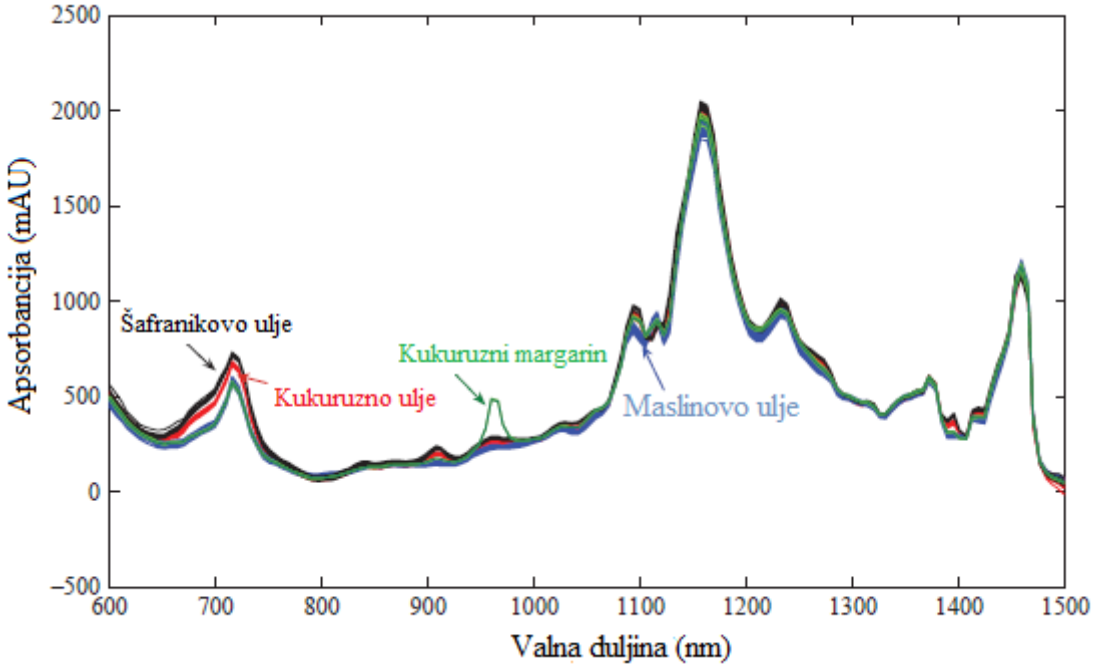
Ovo istraživanje uključuje nastojanje da se uzorci biljnih ulja dodijele jednoj od četiri grupe. Pomoću NIR *(Near Infrared*) spektroskopije, jedne od tradicionalne tehnike za primjenu kemometrije. Budući da je NIR spektroskopija jedna od najranijih tehnika kako bi se izvukla korist od kemometrije, postoji velika količina literature o korištenju različitih metoda. Mnogi pristupi su ipak vrlo specifični za NIR spektroskopiju i nisu primjenjivi na svim područjima prepoznavanja obrazaca. U ovom slučaju, podaci se sastoje od 72 spektra:

1. 18 uzoraka kukuruznih ulja (klasa A)
2. 30 uzoraka maslinovog ulja (klasa B)
3. 16 uzoraka šafranikova ulja (klasa C)
4. 8 uzoraka kukuruznog margarina (klasa D)

Valja imati na umu da je broj uzoraka kukuruznog margarina prilično nizak te se mogu pojaviti problemi u modeliranju grupa s manjim brojem uzoraka. U ovom skupu podataka koriste se sljedeći koraci za pripremu podataka:

1. NIR podaci su korigirani pomoću pristupa zvanog MSC (engl. *Multiplicative Scatter Correction*).
2. Područje spektra između 600 i 1500 nm valne duljine koristi se za prepoznavanje obrazaca.
3. Podaci su prosječni jer su neka područja intenzivnija od drugih, ali je varijabilnost na svakoj valnoj duljini vrlo slična. Nije potrebno standardiziranje.

MSC korigirani spektri prikazani su na slici 8. Može se vidjeti da postoje male razlike između spektara grupa. Na primjer, na oko 700 nm šafranikova ulja vidljive su intenzivnije apsorpcije praćene kukuruznim uljem; međutim te razlike su prilično male i postoji mali razmak između svake klase. Vrlo je teško identificirati nepoznato ulje okom koristeći samo jedan NIR spektar. Tehnike prepoznavanja obrazaca mogu se koristit da se odredi mogu li se razlikovati skupine, koje spektralne značajke su najbolje za diskriminaciju i koliko dobro se nepoznati objekt može dodijeliti specifičnoj klasi.



Slika 8. MSC korigirani NIR spektri od četiri klase ulja[4]

Ova primjena klasična je u kemometriji i jedna je od tehnika prepoznavanja obrazaca koje se uspješno izvode. Na NIR skupovima podataka kemometrijske tehnike obično imaju tendenciju da rade vrlo dobro. [4]

4.2. Slučaj 2: Analiza onečišćenje okoliša pomoću *Headspace* masene spektrometrije

Cilj je proučiti zagađenje detekcijom ugljikovodika u različitim prirodnim staništima, osobito u zonama gdje se sirovo ulje izlučuje i izlijeva. Iako se plinske kromatografske metode tradicionalne koriste u ovakvim slučajevima, dugotrajne su i zahtijevaju poravnanje pikova ili ručno tumačenje. MS (masena spektrometrija) je brža alternativa osim ako se ne poveže sa kromatografijom. Iako možda neće dati informacije o individualnim komponentama u smjesi mogu se koristiti kao "*fingerprint*" da se dobiju informacije obrađene prikladnim kemometrijskim tehnikama s ciljem donošenja odluke.Glavni dodatak za uzorkovanje, u ovom slučaju, dodan je na MS i zajedno su korišteni za analizu tla zagađenog sirovim uljem i derivatima. Skup podataka sastojao se od 213 uzoraka tla i pijeska. Od toga, u

179 uzoraka sadržava ulje i oni predstavljaju zagađene uzorke. Preostalih 34 ostali su "čisti" i predstavljaju nezagađene uzorke (tablica 2). U ovom slučaju primarni problem je određivanje dolazi li uzorak iz zagađene klase (klasa A) ili nezagađene klase (klasa B), više nego razina zagađenja, što je problem kalibracije. Očekujemo slabo zagađene uzorke koji se nalaze blizu granice između klasa. Pokušava se pratiti je li moguće razlikovati zagađeno od nezagađenih uzoraka pomoću MS podataka i prepoznavanja obrazaca. Nakon toga, određuje se koliko dobro možemo klasificirati uzorke u jednu od te dvije klase. MS podaci bilježe karakteristične mase od m/z 49 do 160 (slika 9.). u nekim slučajevima (slika 10.) dominira mali broj velikih pikova. Pripremna obrada podataka mora uzeti u obzir sljedeće korake[4]:

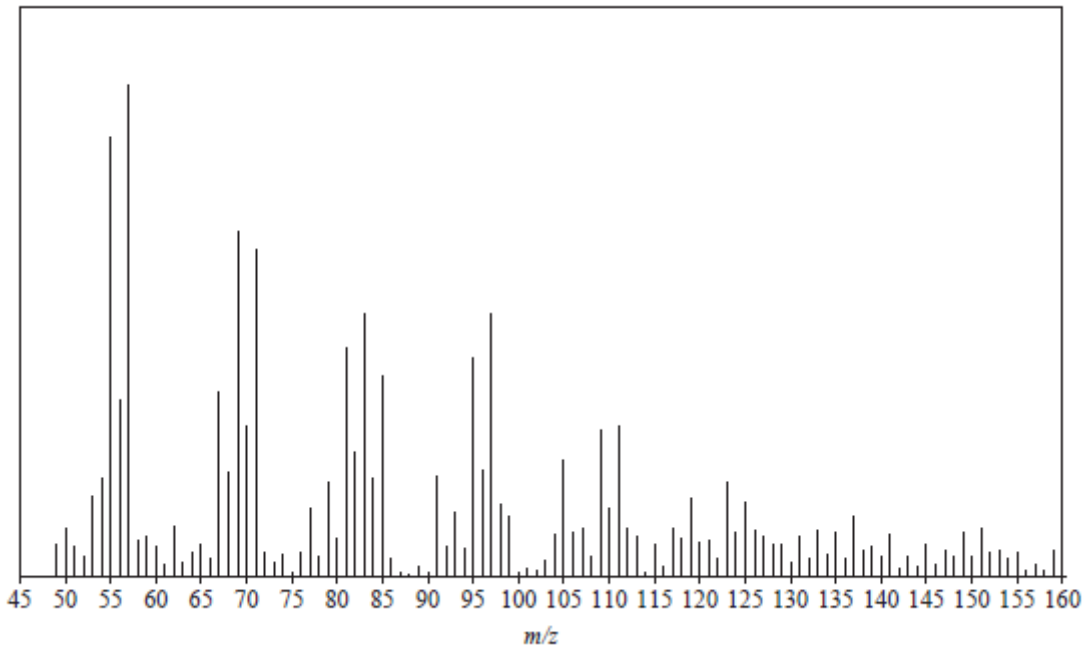
1. Intenzitet MS-a najprije se korjenuje kvadratnim korijenom. To nije uvijek nužno za rukovanje MS podacima, ali ako se taj korak napravi, smanjit će se utjecaj velikih pikova na rezultat.
2. Svaki kvadratni korijen masenog spektra smanjuje se do ukupno jedan zato što količina uzorka uvedenog u MS instrument se ne može lako kontrolirati te je teško pronaći standarde za HS-MS.
3. U konačnici, kolone su standardizirane kako bi se omogućilo da svaka m/z vrijednost ima jednak utjecaj na krajnje prepoznavanje obrazaca. Ponekad se ioni niskog intenziteta mogu dijagnosticirati iz zanimljivih spojeva koji su prisutni u malim količinama.

Tablica 2: Uzorci za slučaj 3[1]

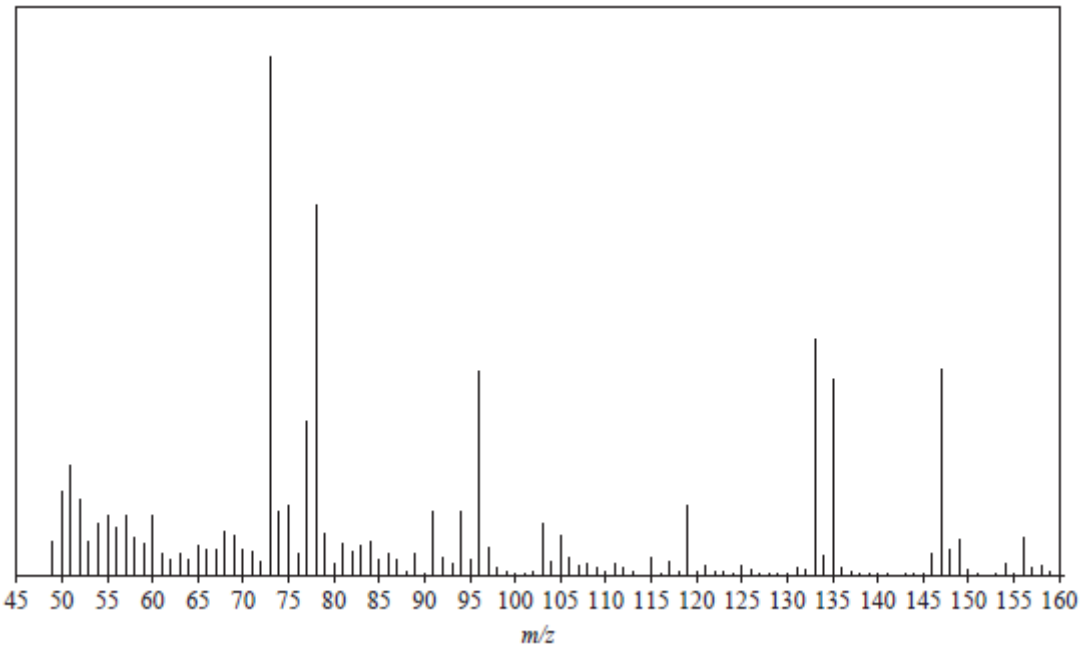
Uzorak	Količina dodanog ulja (mg/kg)	Broj
Zagađeni uzorci (klasa A)	1,2	3
Komercijalni pijesak sa plinskim uljem	2,4	3
	5,9	3
	12	3
	24	3
	59	3
	74	3
	148	3
	297	3
	371	3
Komercijalni pijesak sa Iranskim naftnim uljem	1,4	3
	2,8	3
	6,9	3
	14	3
	28	3
	69	3
	86	3
	172	3
	345	3
	431	3
Komercijalni pijesak sa naftnim uljem iz Nigerije (rijeka Brass)	1,3	3
	2,8	3
	6,9	3
	14	3
	28	3
	69	3
	86	3
	172	3
	345	3
	431	3

Tablica 2: nastavak

Uzorci	Količina dodanog ulja (mg/kg)	Broj
Pijesak sa Santander plaže sa naftnim uljem iz Nigerije (rijeka Brass)	1,3	3
	2,6	3
	6,5	3
	13	3
	26	3
	65	3
	82	3
	164	3
	327	3
	409	3
		3
Nezagađeni uzorci (klasa B)	none	3
Pijesak sa Santander plaže 2	none	3
Pijesak sa Koruna plaže 4	none	3
Tlo 1	none	3
Tlo 2	none	3
Tlo 3	none	3
Tlo 4	none	3
Komercijalni pijesak	none	10
Pijesak sa Santander plaže 1	none	3
Pijesak sa Koruna plaže 3	none	3



Slika 9. Tipični maseni spektar (pijesak sa Santander plaže sa 2,4 mg/kg plinskog ulja) [4]



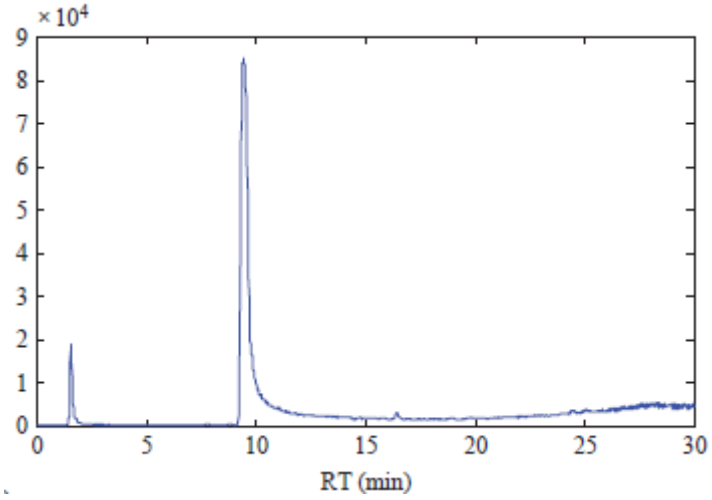
Slika 10. Maseni spektar uzorka tla 3 [1]

4.3. Slučaj 3: Tekućinska kromatografija i masena spektrometrija farmaceutskih tableta

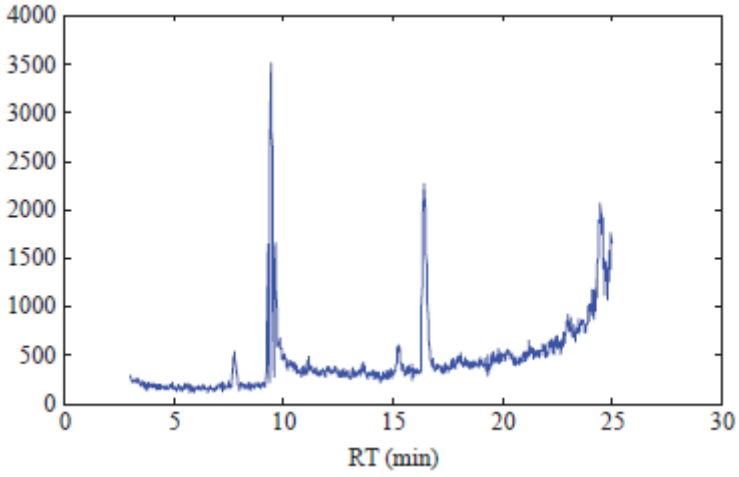
Kemometrika ima glavnu ulogu u farmaceutskoj industriji, posebice s PAT (engl. *Process Analytical Technology*) inicijativom, vezano uz kvalitetu i podrijetlo proizvoda. Za većinu farmaceutskih proizvoda istog lijeka, glavni sastojak, API (engl. *Active Pharmaceutical Ingredient*) je jednak. Male razlike u njihovom sastavu ukazuju na razliku u procesima proizvodnje. Te male nečistoće često dijagnosticiraju specifične probleme u proizvodnji i mogu se koristit kao "fingerprints" podrijetla serije proizvoda, na primjer, iz koje tvornice dolazi, tko je dobavljač, ponaša li se postrojenje slično drugom postrojenju i tako dalje. U LC/MS (engl. *Liquid Chromatography Mass Spectrometry*) nečistoće koje pokazuju znakove kako je, na primjer, tableta proizvedena može se otkriti kao mali pik. Međutim, osim puta proizvodnje, instrument na kojem je uzorak analiziran također utječe na izgled LC/MS zapisa. To je zato što instrumenti (i kromatografske kolone) stare i njihove sposobnosti se mijenjaju; zaprljaju se, čiste se, mijenjaju se kolone i tako dalje. Ne postoje dva instrumenta koja rade jednako. To je realan problem zato što se instrumenti nakon nekog vremena mijenjaju pa će analitička metoda, zbog razvoja laboratorija, razvijena tijekom jedne godine morati biti prebačena na novi instrument iduće godine. To je karakteristično za uređene industrije kao što je farmaceutska industrija gdje postoje konstante promjene u opremi, bilo da zastari, zataji ili se jednostavno istroši. Očekujemo da na LC/MS signal utječe i podrijetlo tablete i instrumentalni uvjeti. U ovom slučaju traži se utvrđivanje podrijetla tableta koje dolaze iz tri proizvodna puta i postoji li specifična oznaka koja se može pronaći u nečistoćama. Nametnut je instrumentalni signal i dva primarna čimbenika utječu na izgled LC/MS signala, put i instrumentalne uvjete. Uzorci su podijeljeni na podrijetla (različiti proizvodni put) i šarže (različiti instrumentalni uvjeti). To je od interesa kako bi se utvrdilo je li moguće klasificirati uzorke prema njihovom podrijetlu (što je primarni cilj ovog istraživanja) ali i koliko su važni instrumentalni uvjeti i koja su obilježja primarna zbog podrijetla, a koja zbog instrumenta. Neobrađeni LC/MS zapis (slika 11.) ne izgleda obećavajuće jer su glavni vrhovi pika pomoćne supstance i API. Kako bi se dobio više informativan zapis napravilo se sljedeće:

1. Područja ispod 3. minute, gdje pomoćna supstanca eluira, i iznad 25. minute su uklonjena.

2. Skidanje pomoćne supstance korišteno je za uklanjanje velikog pika zato što API eluira na otprilike 10 minuta; ovaj postupak omogućava zadržavanje ko-eluenata i alternativa je jednostavnom uklanjanju cijelog područja i velikih pikova, ali način zadržavanja manjih pikova ispod koji pokazuju različite spektre.
3. podaci su obrađeni postupkom CODA koji rezultira da kromatogrami imaju manje šumova. Krajnji LC/MS zapis vidljiv je na slici 12.
4. Pikovi su detektirani u svakom LC/MS zapisu.
5. Pikovi iz svakog LC/MS zapisa koji su imali slično kemijsko podrijetlo bazirano na kromatografskom retencijskom vremenu i sličnoj masenoj spektrometriji su identificirani (90 pikova). Integrirani intenzitet za sve mase je izračunati za svaki pik
6. Konačno, podaci se predstavljaju kao pik tablica dimenzija 79 × 90 gdje se stupci odnose na jedinstvene spojeve, a redovi na uzorke. [4]



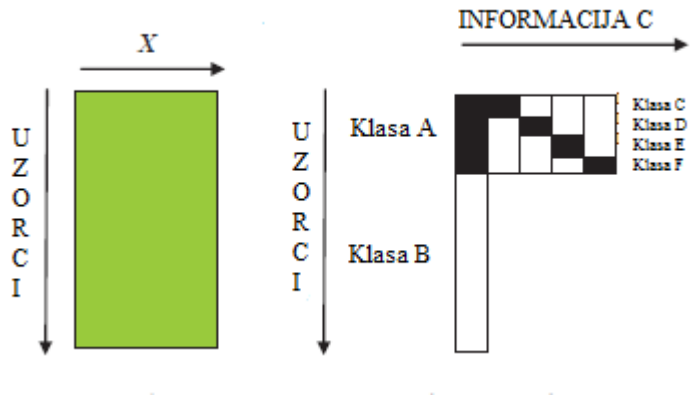
Slika 11. LC/MS zapis tablete; neobrađeni podaci [4]



Slika 12. LC/MS zapis tablete; nakon što su uklonjeni zadržani pikovi između 3. i 25. minute[4]

4.4. Slučaj 4: Atomska spektroskopija za istraživanje hipertenzije

Ovaj skup podataka sadrži 540 uzoraka koji se mogu razdvojiti u 94 uzorka pacijenata (klasa A) koji imaju bolest hipertenzije (visok krvi tlak) i 446 uzoraka kontroliranih pacijenata (klasa B) za koje se zna da nemaju hipertenziju. Klasa A može se podijeliti na 4 različite podklase, klasa C – kardiovaskularna hipertenzija (CV - 31 uzorak), klasa D – kardiovaskularna nesreća (CA - 19 uzoraka), skupina E – bubrežna hipertenzija (RH – 21 uzorak) i klasa F (maligna hipertenzija (MH – 23 uzorka). Struktura ovog skupa podataka je djelomično hijerarhijska (slika13.).



Slika 13. Struktura slučaja [4]

U skupu podataka postoji samo pet varijabli, a to su količine bakra (Cu), cinka (Zn), magnezija (Mg), olova (Pb) i kadmija (Cd) u krvi u jedinicama od $\mu g/ml$. Sve su otkrivene na različitim rasponima (tablica 3) pa su podaci standardizirani prije prepoznavanja obrazaca.

Tablica 3: Raspon u kojem je pet elemenata pronađeno u krvi ($\mu g/ml$) [1]

Element	**Cu**	**Zn**	**Mg**	**Pb**	**Cd**
Maksimum	1,8	13,2	79,5	321	171,1
Minimum	0,42	4,9	14,5	0,078	0,08

Postoje neke značajke ovog skupa podataka:

1. Postoji vrlo malo varijabli.
2. Veličine klasa A i B su vrlo različite i klasa A ima podklase.
3. Neke sličnosti (između kontroliranih i bolesnih pacijenata) vidljive su okom, a mogu se odrediti koristeći jednostrano mjerenje (tablica 4).

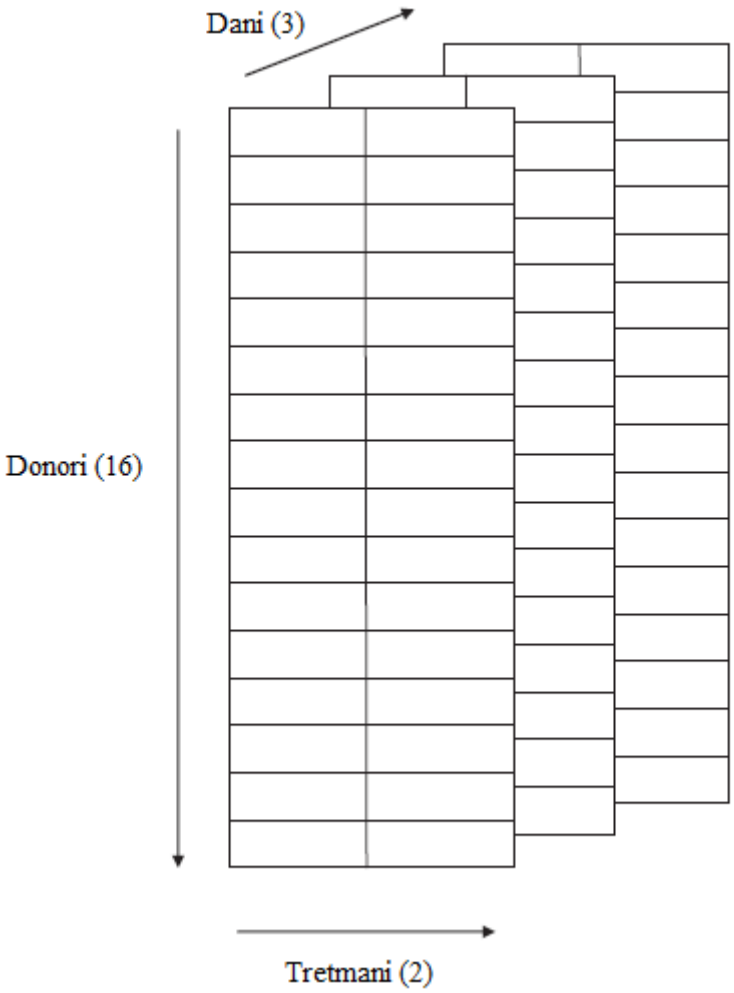
Ovaj skup podataka je vrlo jednostavan u usporedbi s drugim primjenama.[4]

Tablica 4: Rasponi pet elemenata u slučaju (μg/ml) [1]

Element	Cu	Zn	Mg	Pb	Cd		Cu	Zn	Mg	Pb	Cd
A	1,6	9,3	52,1	0,8	9,9						
	0,06	4,9	14,5	0,078	0,08						
B	1,8	13,2	79,5	321	171,1	**C**	1,8	9,9	79,5	321	132
	0,42	6,1	32,4	131	60		0,62	6,3	32,4	195	81,3
						D	0,89	10,2	56,9	307,1	98,3
							0,42	6,5	36,2	165,7	60
						E	0,91	13,2	51,2	314	171,1
							0,57	6,9	36,5	198	102,9
						F	1,7	11,7	57,5	297	106
							0,59	6,1	37,5	131	79

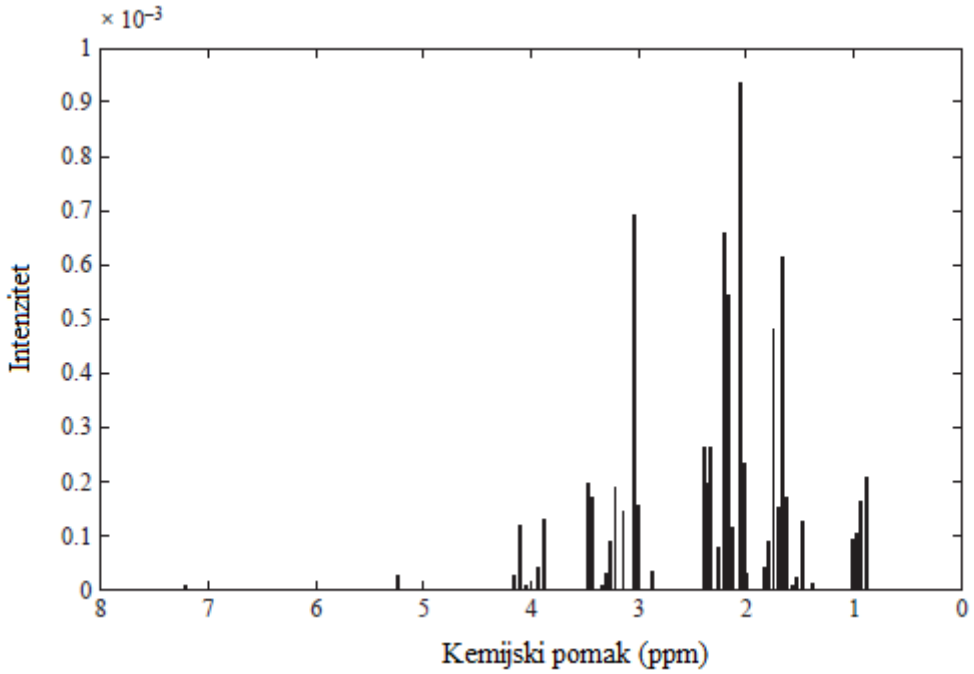
4.5. Slučaj 5: Nuklearna magnetska rezonancijska spektroskopija (engl. *Nuclear Magnetic Resonance Spectroscopy*) za analizu sline efektom ispiranja usta

NMR spektroskopija koristi se kako bi se odredio efekt ispiranja usta na uzorcima sline. Volonteri (16 donora) koji su odabrani nemaju niti jedan oblik aktivne zubne bolesti. Svaki donor dao je uzorak sline nakon jutarnjeg buđenja. Svaki uzorak je podijeljen u dva dijela i tretiran sa 0.5 ml oralne formulacije za ispiranje usta (klasa A) ili vodom (klasa B) i svaki donor je proučavan tijekom tri dana s jednim uzorkom po danu. Struktura skupova podataka prikazana je na slici 14. 16 donora × 2 tretmana × 3 dana = 96 uzoraka.



Slika 14. Struktura slučaja 6 [4]

^{1}H NMR spektroskopija je korištena za analizu svakog uzorka i podacima podijeljenima u 146 pametnih jedinica (engl. *intelligent buckets*) što je način podjele spektra izvan određenih ppm regija. Intenzitet svake jedinice se sažima u svakom spektru iznad odgovarajuće regije. Razlog toga, a ne korištenje neobrađenih NMR podataka je postojanje malih pomaka između spektara. Te male promjene u prosjeku smanjuju pogreške koje bi se inače dogodile. Neka područja spektra koja su bez informacija uklanjaju se tim postupkom; sve jedinice koje imaju intenzitet manji od 1% se također uklanjaju zbog toga što najčešće sadržavaju šumove. Osim nekoliko malih jedinica u aromatskom području, sve kemijske promjene su bile manje od 5,44 ppm. Preostaje 49 jedinica. Na slici 15. prikazan je tipični NMR spektar i stvorena je matrica podataka 96 × 49 za naknadno prepoznavanje obrazaca. [4]



Slika 15. Tipični NMR spektar korišten u slučaju 6 [4]

5. ZAKLJUČAK

Korištenje prepoznavanja obrazaca i klasifikacije od temeljne je važnosti za mnoge automatizirane elektroničke sustave koji se danas koriste. Njihove primjene su širokog spektra, od analitičke kemije, medicinske dijagnoze do vojne obrane, od biometrije do strojnog učenja i tako dalje. Prepoznavanje obrazaca značajno se razvilo tijekom šezdesetih godina dvadesetog stoljeća, no i dalje je vidljivo kako postoji prostor za napredak i razvoj. To je pristup koji se koristi za smanjenje podataka koji ne daju relevantne informacije. U današnje vrijeme ima mnogo više kemijskih informacija nego što ih se može obraditi. Može biti vrijedan dio tumačenja kemijskih podataka pomoću računala. Bez obzira što postoji velik broj literature, ova tema i dalje ostaje vrlo izazovna i zahtjeva, posebice za nekoga koji nije stručnjak u tom području. Tijekom proteklih desetljeća dolazi do uključivanja novih izvora podataka, a to posebno zbunjuje mnoge analitičke kemičare. Razvoj, primjerice korištenje kombinirane kromatografije, masene spektrometrije, NMR spektroskopije bio je vrlo brz, s poboljšanim, osjetljivim i automatiziranim instrumentima. Naizgled, sve izgleda vrlo jednostavno, ali nije. Dolazi se do problema da skupovi podataka postaju znatno teži za rukovanje, te više nisu jednostavni NIR spektri koje su kemometričari lako svladavali. Potencijalna promjena kemometrije na analitičke podatke koji proizlaze iz problema u biologiji, medicini je ogromna, ali često znanstvenici nemaju razumijevanja kako stjecati podatke i njima rukovati. Većina kemometrijskih metoda ima svoje podrijetlo u analitičkoj kemiji, gdje često postoje jasne činjenice. Na primjer, u kalibraciji se zna koji se odgovor traži te se multrivarijantne metode pokušavaju dovesti što bliže poznatom odgovoru. U nekim od izvornih primjena kemijskog poznavanja uzorka kao što je spektroskopija zna se što su temeljne skupine spojeva i želi se klasificirati spektre što je moguće učinkovitije u tim skupinama. Cilj je stopostotna točnost i algoritmi se smatraju boljima što je odgovor točniji.

6. LITERATURA

1. Dougherty G., Pattern Recognition and Classification, Springer, Kalifonija, SAD, 2013.
2. Vercellis C., Business Intelligence: Data Mining and Optimization for Decision Making, John Wiley & Sons Ltd., Milano, Italija, 2009.
3. Witten I. H., Frank E., Data Mining; Practical Machine Learning Tools and Techniques, 2 izd, Morgan Kaufmann Publishers, San Francisco, Kalifornija, 2005.
4. Brereton, R. G., Chemometrics for Pattern Recognition, John Wiley & Sons Ltd., Bristol, Velika Britanija, 2009.
5. Varmuza K., Pattern Recognition in Analytical Chemistry, Analytica Chimico Acta. **122** (1980) 227-240
6. Brereton, R. G., Chemometrics Data Analysis for the Laboratory and Chemical Plant, John Wiley & Sons Ltd, Bristol, Velika Britanija, 2003.
7. https://ch.mathworks.com/discovery/machine-learning.html, pristupljeno 25. 8. 2018.
8. Perović V., Primjena statističkih modela u predviđanju nalazišta biljnih vrsta, diplomski rad, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Zagreb, 2016.
9. Everitt B. S., Landau S., Leese M., Stahl D., Cluster Analysis, 5. izd, John Wiley & Sons Ltd, London, Velika Britanija, 2011.
10. Ukić Š., Nazivlje u kemometriji/kemometrici?, Imenje i nazivlje, Kemija u industriji, **65** (3-4) (2016) 181–182
11. Perner P., Petrou M., Machine Learning and Data Mining, Springer, Leipzig, Njemačka, 1999.
12. Einax J., Chemometrics in Environmental Chemistry, Springer, Berlin, Njemačka, 1995.
13. Theodoridis S., Koutroumbas K., Pattern Recognitions, 5. izd., Elsevier, London, Velika Britanija, 2009.
14. Duda R. O., Hart P. E., Stork D. G., Pattern Classification, 2. izd., John Wiley & Sons, Inc., Londin, Velika Britanija, 2001.
15. Kowalski B. R., Chemometrics; Mathematics and Statistics in Chemistry, D. Reidel Publishing Company, Dordrecht, Nizozemska, 1984.
16. Adams M. J., Chemometrics in Analytical Spectroscopy, 2. izd., The Royal Society of Chemistry, London, Velika Britanija, 2004.

17. Brown S. D., Tauler R., Walczak B., Comprehensive Chemometrics; Chemical and Biochemical Data Analysis, Elsevier, London, Velika Britanija, 2009.
18. Brereton R. G., Applied Chemometrics for Scientists, John Wiley & Sons Ltd, London, Velika Britanija, 2007.
19. Marini F., Chemometrics in Food Chemistry, Elsevier, Oxford, Velika Britanija, 2013.
20. Langley P., Elements of Machine Learining, Morgan Kaufmann Publishers, Inc., San Francisco, Kalifornija, SAD, 1996.
21. Abonyi J., Feil B., Cluster Analysis for Dana Mining and System Identification, Birkhauser Verlag AG, Basel, Švicarska, 2007.
22. Romesburg H. C., Cluster Analysis for Researchers, Lulu Press, Sjeverna Karolina, SAD, 2004.
23. Hand D., Mannila H., Smyth P., Principles od Dana Mining, The MIT Press, London, Velika Britanija, 2001.
24. Houck M. M., Forensic Chemistry, Elsevier, Oxford, Velika Britanija, 2015.
25. Smilde A., Bro R., Geladi P., Multi-way Analysis Applications in the Chemica Sciences, John Wiley & Sons, Ltd, London, Velika Britanija, 2004.
26. Workman J. Jr., Springsteen A., Applied Spectroscopy; A Compact Reference for Practitioners, Academic Press, San Diego, Kalifornija, SAD, 1998.
27. Belton P. S., Engelsen S. B., Jakobsen H. J., Magnetic Resonance in Food Science, The Royal Society of Chemistry, Cambridge, Velika Britanija, 2005.
28. Workman J. Jr., Mark H., Chemometrics in Spectroscopy, Academic Press, San Diego, Kalifornija, SAD, 2007.
29. Mannhold R., Krogsgaard – Larsen P., Timmerman H., Chemometrics Methods in Molecular Design, VCH Verlagsgesellschaft, Weinheim, Njemačka, 1995.
30. Otto M., Chemometrics, Statistics and Computer Aplication in Analytical Chemistry, 3. izd, Wiley-VCH Verlag GmbH & Co., Weinheim, Njemačka, 2017.
31. Clarke B., Fokoue E., Zhang H. H., Principles and Theory for Data Mining and Machine Learning, Springer, Heidelberg, Njemačka, 2009.
32. Chen H., Fuller S. S., Friedman C., Hersh W., Medical Informatics: Knowledge Management and Data Mining in Biomedicine, Springer, New York, SAD, 2005.
33. Giudici P., Applied data mining: statistical methods for business and industry, John Wiley & Sons, Ltd, London, Ujedinjeno Kraljevstvo, 2003.
34. Berrueta L. A, Alonso-Salces R. A., H´eberger K., Supervised pattern recognition in food analysis, Journal of Chromatography A **1158** (2007) 196–214

35. Fernandesa A. P., Santosa M. C., Lemosa S. G., Ferreirac M. M. C., Nogueirab A. R. A., No´brega J. A., Pattern recognition applied to mineral characterization of Brazilian coffees and sugar-cane spirits, Spectrochimica Acta Part B **60** (2005) 717– 724
36. Gazzaz N. M., Yusoff M. K., Ramli M. F., Aris A. Z., Juahir H., Characterization of spatial patterns in river water quality using chemometric pattern recognition techniques, Marine Pollution Bulletin **64** (2012) 688–698
37. Coomans D., Massart D. L., Alternative k-nearest neighbour rules in supervised pattern recognition, Analytica Chimica Acta. **136** (1982) 15-27
38. Metzlaff M., O'Dell M., Cluster P. D., Flavell R. B., RNA-Mediated RNA Degradation and Chalcone Synthase A Silencing in Petunia, Cell, **88** (1997), 845–854
39. Bro R., Multivariate calibration; What is in chemometrics for the analytical chemist?, Analytica Chimica Acta **500** (2003) 185–194
40. Belton P. S., Colquhoun I. A., Kemsley E. K., Delgadillo I., Roma P., Dennis M. J., Sharman M., Holmes E., Nicholsond J. K., Spraule M., Application of chemometrics to the 1H NMR spectra of apple juices: discrimination between apple varieties, Food Chemistry **61** (1998) 207-213
41. Davatzikos C., Ruparel K., Fan Y., Shen D. G., Acharyya M., Loughead J. W., Gur R. C., Langlebenb D. D., Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection, NeuroImage **28** (2005) 663 – 668
42. Hagberg G., From magnetic resonance spectroscopy to classification of tumors. A review of pattern recognition methods, NMR in Biomedicine **11** (1998), 148–156
43. Wei L., Yang Y., Nishikawa R. M., Jiang Y., A Study on Several Machine-Learning Methods for Classification of Malignant and Benign Clustered Microcalcifications, Iee Transactions on Medical Imaging, **24** (3) (2018) 371-380
44. Galas D. J., Eggert M., Waterman M. S., Rigorous Pattern-recognition Methods for DNA Sequences; Analysis of Promoter Sequences from Escherichia coli, Journal of Molecular Biology **186** (1985) 117-128
45. Webb A. R., Copsey K. D., Statistical Pattern Recognition, 3. izd, John Wiley & Sons Ltd, London, Velika Britanija, 2011.
46. Elloumi M., Iliopoulos C., Wang J. T. L., Zomaya A. Y., Pattern recognition in Computational Molecular Biology: Techniques and Approaches, John Wiley & Sons Ltd, London, Velika Britanija, 2016.
47. Miller J. N., Miller J. C., Statistics and Chemometrics for Analytical Chemistry, 5. izd., Pearson Education Limited, Engleska, Ujedinjeno Kraljevstvo 2005.

48. Seber G. A. F., Lee A. J., Linear Regression Analysis, John Wiley & Sons Ltd., London, Velika Britanija, 2003.
49. Montgomery D. C., Peck E. A., Vining G. G., Introduction to Linear Regresion Analysis, John Wiley & Sons Ltd., London, Velika Britanija, 2006.
50. Metclaf G. S., Windig W., Hill G., Meuzelaar H. L. C., Characterization of U.S. Lignites by Pyrolysis; Mass Spectrometry and Multivariate Analysis, International Journal of Coal Geology, **7** (1987) 245-268
51. Kowalski B. R., Bender C. F., The K-Nearest Neighbor Classification Rule(Pattern Recognition) Applied to Nuclear Magnetic Resonance Spectral Interpretation, Analytical Chemistry, **44**, (1972), 1405-1411
52. Hilario M., Kalousis A., Pellegrini C., Muller M., Processing and Classification of Protein Mass Spectra, Mass Spectrometry Reviews, **25** (2006) 409-449

ŽIVOTOPIS

Anja Rakas [redacted] Osnovnu školu pohađala je u Glini. Srednjoškolsko obrazovanje stekla je u Gimnaziji Sisak, smjer- Opća gimnazija. 2014. godine upisuje studij Primijenjena kemija na Fakultetu kemijskog inženjerstva i tehnologije.

Tijekom studija odradila je stručnu praksu u farmaceutskoj kompaniji Pliva d.o.o na odjelu Istraživanje i razvoj, Kemija TAPI, R&D.