# Tehnike otkrivanja kontrasta

**Bali, Helena**

**Undergraduate thesis / Završni rad**

**2020**

*Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj:* **University of Zagreb, Faculty of Organization and Informatics / Sveučilište u Zagrebu, Fakultet organizacije i informatike**

*Permanent link / Trajna poveznica:* https://urn.nsk.hr/urn:nbn:hr:211:951617

**SVEUČILIŠTE U ZAGREBU**

**FAKULTET ORGANIZACIJE I INFORMATIKE**

**V A R A Ž D I N**

**Helena Bali**

# TEHNIKE OTKRIVANJA KONTRASTA

**ZAVRŠNI RAD**

**Varaždin, 2020.**

**SVEUČILIŠTE U ZAGREBU**

**FAKULTET ORGANIZACIJE I INFORMATIKE**

**V A R A Ž D I N**

**Helena Bali**

**Matični broj: 44070/15–R**

**Studij: Poslovni sustavi**

# TEHNIKE OTKRIVANJA KONTRASTA

**ZAVRŠNI RAD**

**Mentorica:**

Doc. dr. sc. Dijana Oreški

**Varaždin, lipanj 2020.**

*Helena Bali*

**Izjava o izvornosti**

Izjavljujem da je moj završni rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristila drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

*Autorica potvrdila prihvaćanjem odredbi u sustavu FOI-radovi*

_____

**Sažetak**

Tehnike otkrivanja kontrasta (eng. *contrast data mining techniques*) jedne su od vrsta klasifikacijskih tehnika pomoću kojih se može „rudariti“ odnosno filtrirati specifično znanje iz većih skupova podataka. Kontrast u ovom kontekstu zapravo znači promijena, razlika ili usporedba, stoga se spomenute tehnike koriste kako bi se uočile razlike između danih klasa objekata.

STUCCO i Magnum Opus algoritmi, tehnike su otkrivanja kontrasta i u ovom su završnom radu glavna tema teoretskog pregleda. Nakon teoretskog pokrivanja, uslijedit će usporedba algoritama. Osim toga, biti će navedeno gdje su se spomenute tehnike uspješno primjenjivale do sada uz detaljan opis jedne primjene i izrade vlastitog primjera. U samom zaključku, autorica će iznijeti svoje skromno mišljenje u vezi rudarenja podataka općenito te će zaključiti što je potrebno zapamtiti u vezi promatranih algoritama.

**Ključne riječi:** rudarenje podataka, *data mining*, tehnike otkrivanja kontrasta, *contrast data mining techniques*, STUCCO algoritam, Magnum Opus algoritam

# Sadržaj

# 1. Uvod

Jedna od najzanimljivijih tehnologija informacijskih znanosti u dvadesetprvom stoljeću jest područje rudarenje podataka. Rečenicu izjavljuju Dong i Bailey (2012), a iz nje proizlazi motivacija autorice završnog rada za odabir teme.

Rudarenje podataka (eng. *data mining*) organiziranje je, sortiranje ili grupiranje velikog broja podataka te izvlačenje relevantnih informacija pomoću specifičnih tehnika. Nestrukturirani skupovi podataka sačinjavaju prema nekim izvorima devedeset posto digitalnog svijeta pa se postavlja pitanje: „*Zašto je rudarenje podataka važno?*“ Odgovor je sljedeći: omogućuje snalaženje u kaosu neprovjerenih i ponovljivih informacija, pomaže u razumijevanju što uistinu jest važno tako da se određene informacije koriste u najprikladnim situacijama, ubrzava ritam donošenja potencijalno ispravnijih odluka. (*SAS Institute*, bez dat.)

Rudarenje podataka primjenjuje se međuostalom u komunikacijske, obrazovne ili bankarske svrhe. Multimedijske i telekomunikacijske tvrtke koriste analitičke modele kako bi iz mora korisničkih informacija (eng. *customers data*), predvidjeli korisničko ponašanje te svojim korisnicima ponudili specifičnije i relevantnije ponude namjenjene isključivo njima. S unificiranim, podatkovno potaknutim (eng. *data-driven*) pogledom na učenikov napredak, profesori mogu predvidjeti učeničke performanse bez kročenja u učionicu i samim time mogu razviti strategije  koje omogućuju učenicima da, na primjer, ne zaostaju. Rudarenje podataka može skrenut pozornost na učenika koji možda treba dodatnu pomoć. Automatizirani algoritmi omogućuju bankarima bolji uvid u tržišne rizike, brže uočavanje prevara i dobivanje optimalnog povrata sredstava u njihova investiranja. (*SAS Institute*, bez dat.)

Definiranje pojma rudarenja podataka, nadovezuje se na pojam kontrastnog rudarenja podataka (eng. *contrast data mining*). Shah i Dinkal (2015)  kažu da kontrastno rudarenje podataka uključuje usporedbu jednog skupa, vrste ili klase objekata s drugim skupom, vrstom, klasom objekata. Koristi se kako bi se uočili razlike između danih objekata. Dobivene informacije informiraju o onome što razlikuje jedan skup podataka od drugog skupa podataka. Razumijevanje po čemu se skupovi podataka razlikuju, pomaže kako da se određeni objekti koriste u prikladnim situacijama, na primjer za povećavanje profita, smanjivanje troškova i rizika.

Struktura rada sadrži uvodni dio, primjenu korištenih metodika i tehnika pisanja rada, detaljan teoretski opis tehnika otkrivanja kontrasta, usporedbu algoritama STUCCO i Magnum Opus, njihovu primjenu te vlastiti primjer primjene autorice.

# 2. Tehnike otkrivanja kontrasta

Informacija su moć te one pokreću svijet. Veća je potreba za znanjem danas, nego što je bila jučer pošto se svijet mijenja brže i sve je teže pratiti promijene tržišta, korisnika, inovacija i slično.

Boettcher (2011) primjećuje slijedeće:

> *„…the key to survival for businesses is the ability to detect, assess, and respond to changing conditions timely and intelligently…“* (Boettcher, 2011)

U prijevodu, ako želite da vaše poslovanje „preživi“, morate biti sposobni inteligentno, pravovremeno otkriti promijene i prilagoditi se istima. Razumijeti promijenu te saznati što više informacija o njoj je ključ.

Tehnikama otkrivanja kontrasta, opisani su različiti tipovi kontrastnih uzoraka. Kontrastno rudarenje podataka, ukratko, traži značajnu i veliku razliku između 2 ili više uzorka/grupa/skupina/klasa, a što više razlika se pronađe, to ste bogatiji informacijama koje možete iskoristiti.

U narednim poglavljima koja dolaze, slijedi upoznavanje s tehnikama otkrivanja kontrasta: STUCCO i Magnum Opus algoritmima, koji pomažu u boljem shvaćanju promijena za kvalitetnije „hvatanje“ znanja.

## 2.1 STUCCO algoritam

STUCCO (*Search and Testing for Understandable Consistent Contrast*) algoritam je koji je prvi puta razvijen u svrhe pronalaženja razlikovnih skupova koji predstavljaju „…conjuctions of attributes and values that differ meaningfully in their distribution across groups…“ (Bay i Pazzani, 2001) ili kako Oreški (2014) prevodi: „..vezu atribut-vrijednost koja značajno razlikuje distribucije među grupama…“ (Oreški, 2014).

Algoritam radi po principu pretraživanja stabla sa svim mogućim razlikovnim skupovima (eng. *contrast set*) gdje u korijenu postoji prazan razlikovni skup. Djeca se stvaraju dodavanjem dodatnih uvjeta koji prate postojeće uvjete u određenom redoslijedu te tako nastaju kombinacije atributa. Formirano stablo, pretražuje se tako što se prvo provjeravaju pojedinačni atributi, a zatim kombinacije atributa. Skup podataka se na svakoj razini pretražuje i svakom razlikovnom skupu se broji podrška (eng. *support*). Svaki čvor se potom pretražuje da bi se utvrdilo da li je *Značajan* i *Velik* (kasnije će biti objašnjeno detaljnije), da li bi ga trebalo „obrezati“ (eng. *prune*) i da li se možda trebaju generirati nova djeca. Kada su svi značajni

razlikovni skupovi locirani, postproces izabire podskup da pokaže korisniku: niži poredak-manje značajne rezultati pa slijede značajniji rezultati koji su iznenađujući i različiti („*Contrast set learning*“, bez dat.).

Podrška dolazi od testiranja nulte hipoteze kojoj je razlikovni skup podrške jednak u svim grupama. Podrška za svaku grupu učestala je vrijednost koja se može analizirati u tablicama nepredviđenih slučaja gdje svaki red predstavlja vrijednost istine razlikovnog skupa, a svaki stupac reprezentira učestalost pripadnosti grupi. Ako postoji razlika u proporcijama između učestalosti razlikovnih skupova i onih od nulte hipoteze, algoritam mora odlučiti, da li ta razlika među proporcijama predstavlja povezanost između varijabli ili može biti pripisana nekom slučajnom uzroku („*Contrast set learning*“, bez dat.).

Dakle, cilj je pronaći razlikovne skupove čija se podrška značajno razlikuje među grupama. Ti razlikovni skupovi, kako Arango (bez dat.) govori u svojem djelu, nazivaju se odstupanja (eng. *deviations*). Sada nešto o odstupanju. Ponovo, Arango (bez dat.) navodi da je odstupanje razlikovni skup koji je *Značajan* i *Velik*. Razlikovni skup u kojem se barem 2 grupe razlikuju u podršci naziva se *Značajan* (eng. *Significant*). Razlikovni skup u kojem je maksimalna razlika između podrške veća od parametra *mindev* naziva se *Velik* (eng. *Large*) (Arango, bez dat.).

Sada će se gore navedene definicije i izraze primjeniti na primjeru. Primjer je preveden i preuzet iz Aranginog djela „*Mining for Contrasting Sets (STUCCO)*“.

Postavljeno je slijedeće pitanje: *Kako se perspektivni studenti iz različitih odsjeka razlikuju jedni od drugih?*

Postoje skupovi studenta s odsjeka računalnih znanosti, biologije i inženjerstva. Nakon što je to definirano, postavlja se podatkovni model (eng. *data model*). To je skup *k*-dimenzionalnih vektora gdje svaka komponenta može sadržavati konačan broj diskretnih vrijednosti (skup mogućih vrijednosti je prebrojiv). Prikažimo potrebne informacije u tablici gdje je *k=6*:

Tablica 1. Podatkovni model

| **Godine** | **Spol** | **Rođen/ a u SAD-u** | **SAT-M > 700** | **SAT-V > 700** | **Upisao/ la** |
|---|---|---|---|---|---|
| < 20 | M | Da | Da | Da | Da |
| 20-25 | M | Da | Ne | Da | Ne |
| 25-30 | F | Ne | Da | Ne | Da |
| … | … | … | … | … | … |

(Prema: Arango, bez dat.)

Tablica 1 prikazuje podatkovni model i neku podjelu po atributima u stupcima: godini, spolu, da li je student/ica rođen/a u SAD, da li je student/ica na SAT ispitu za upis na fakultet imao/la iz matematike više od 700 bodova i isto tako iz jezičnog SAT ispita te da li je osoba upisana u određeni odsjek. Raspon godina je {<20, 20-25, 25-30, >30}, spol može biti muški (M) ili ženski (F), odgovor na pitanje da li je student rođen u SAD može biti da ili ne, isto tako vrijedi i za preostale atribute (Arango, bez dat.)

Vektore organiziramo u međusobno isključive grupe (tablice 2, 3 i 4 u nastavku):

Tablica 2. Grupa- *Računalne znanosti*

| Godine | Spol | Rođen/a u SAD-u | SAT-M > 700 | SAT-V > 700 | Upisao/la |
|---|---|---|---|---|---|
| < 20 | F | Da | Da | Ne | Da |
| 20-25 | M | Ne | Ne | Da | Ne |
| < 20 | F | Ne | Da | .Da | Da |

(Prema: Arango, dez dat.)

Tablica 3. Grupa- *Biologija*

| Godine | Spol | Rođen/a u SAD-u | SAT-M > 700 | SAT-V > 700 | Upisao/la |
|---|---|---|---|---|---|
| 20-25 | M | Da | Da | Ne | Da |
| < 20 | F | Da | Ne | Da | Ne |
| < 20 | F | Ne | Da | .Ne | Da |

(Prema: Arango, dez dat.)

Tablica 4. Grupa- *Inženjerstvo*

| Godine | Spol | Rođen/a u SAD-u | SAT-M > 700 | SAT-V > 700 | Upisao/la |
|---|---|---|---|---|---|
| < 20 | M | Da | Ne | Da | Da |
| 20-25 | M | Da | Ne | Ne | Ne |
| 25-30 | F | Da | Da | .Ne | Da |
| < 20 | F | Da | Ne | Da | Da |

(Prema: Arango, dez dat.)

Primjećuje se slijedeće: postoji netko iz odsjeka biologije čije su godine u rasponu od 20 do 25, muško je, rođen je u SAD-u, kaže da je na matematičkom SAT ispitu imao više od 700

bodova, a na jezičnom SAT manje od 700 bodova i da je upisan u odsjek biologije (Arango, bez dat.).

Razlike među grupama se izražavaju kao razlikovni skupovi (eng. *contrast sets*), a razlikovni skup je par veza atribut-vrijednost, kao što je definirano ranije.

Primjer veze atribut-vrijednost je: *A ∧ B ∧ C: (Godine= 20-25) ∧ (Upisao/ la = Da) ∧ (SAT-V > 700 = Ne)*

Znak ∧ operator je koji ukazuje na logičku konjukciju. *A ∧ B* su istinti ako i samo ako je *A* istinit i ako je *B* istinit. Ovo pomaže u određivanju podrške za razlikovni skup.

Treba odrediti podršku za svaki razlikovni skup. Podrška će dati određeni postotak u svakom odsjeku gdje je razlikovni skup istinit (eng. *true*).

Potrebno je pogledati tablicu 2 i primjeniti izraz za određivanje podrške na tablici za *računalne znanosti*:

*podrška (Spol= F ∧ Rođena u SAD-u = Ne | Računalne znanosti)*

Pogledom na tablicu 2, vidimo da je taj izraz istinit samo u zadnjem (3.) retku te tablice pa je postotak 1/3 odnosno 33 %.

Isto tako se dobiva za izraz u tablici 3:

*podrška (Spol = F ∧ Rođena u SAD-u = Ne | Biologija) = 2 / 3 = 66%*

Potom i za zadnji odsjek inženjerstva, pomoću tablice 4:

*podrška (Spol = F ∧ Rođena u SAD-u = Ne | Inženjerstvo) = 0 / 3 = 0%*

Vidimo da u tablici *inženjerstvo* niti jedan red nije istinit.

Treba pronaći razlikovni skup koji razlikuje jednu grupu odnosno odsjek od drugog. Cilj je pronaći značajne razlike među odsječcima. Pošto se traže značajne razlike, automatski je potrebno definirati ranije spomenuta odstupanja (eng. *deviations*) (Arango, bez dat.).

Primjer izračuna podrške za razlikovni (kontrastni) skup *c1* i postotak parametra *mindev* je *5%*:

*podrška (Upisao/ la = Da ∧ Godine 20-25 | Računalne znanosti) = 11%*

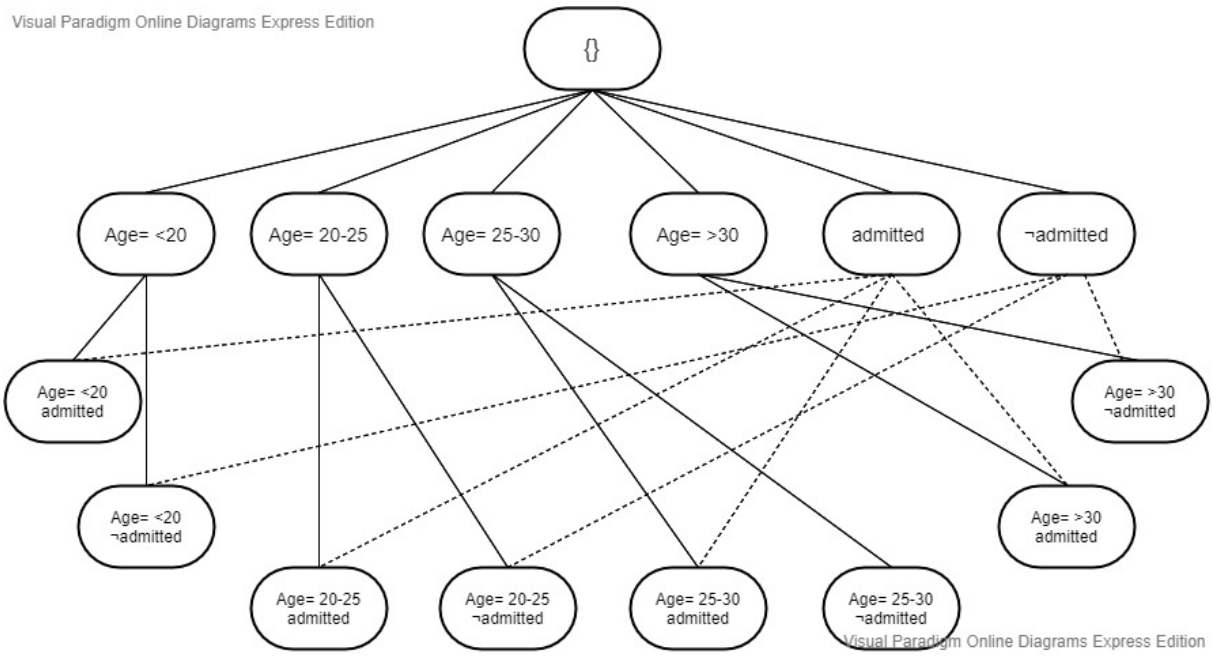
*podrška (Upisao/ la = Da ∧ Godine 20-25 | Biologija) = 15%*

*podrška (Upisao/ la = Da ∧ Godine 20-25 | Inženjerstvo) = 18%*

*Velik* razlikovni skup određuje se tako što se maksimalna razlika između postotaka podrške *(18% -11% = 7%)* usporedi s *mindev* parametrom. Ranije je spomenuto da je razlikovni skup *Velik* ako je maksimalna razlika između postotaka podrške veća od *mindeva.*

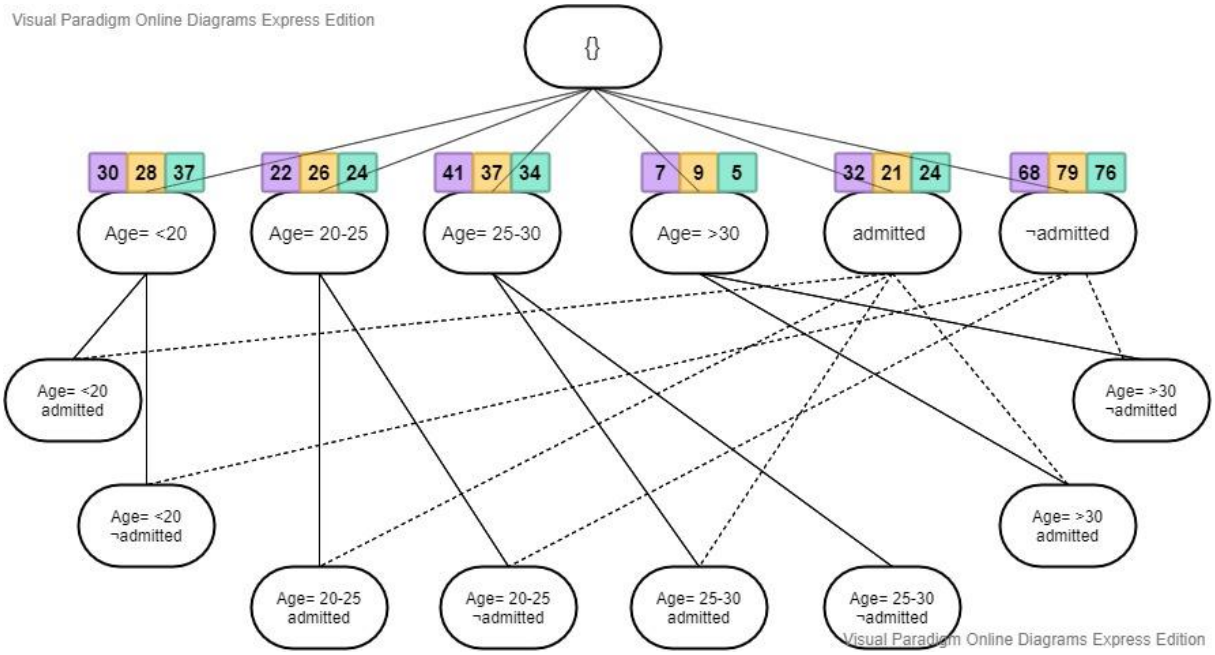
Zaključuje se da je *c1 Velik*. Određivanje da li je razlikovni skup *Značajan*, provodi se pomoću skraćenog statističkog testa kasnije u radu (Arango, bez dat.).

Slijedi prikaz problema modeliranog u stablo. Prvo se pretražuje tako što se gledaju pojedinačni atributi, a zatim kombinacije atributa. To je istaknuto strelicama na slici 1. Gornja strelica prikazuje sve moguće parove veza atribut-vrijednost, a donja strelica vezu između 2 para veze atribut-vrijednost.




Slika 1. Problem modeliran u stablo (Prema: Arango, bez. dat.)

Zatim algoritam pretražuje bazu podataka te vrši izračun podrške za svaku grupu. Ti podaci vidljivi su na slici 2.




Slika 2. Podaci za podršku (Prema: Arango, bez. dat.)

Za svaki čvor treba odrediti da li je *Velik* i *Značajan*. Arango navodi da je razlikovni skup *Značajan* ako se u njemu barem 2 grupe razlikuju po podršci. Izvodi se statistički test „*hi-kvadrat* pod engleskim nazivom „*chi square*“ za razlikovne skupine.

Test uključuje:

- nultu hipotezu koja u prijevodu glasi „Podrška za sve razlikovne skupove je ista u svim grupama.“
- izračun statistike za $\chi^2$
- provjeru vrijednosti *hi-kvadrat* distribucije
- mora biti manji nego od praga *α* (obično je alfa=0.05)

Za izračun *hi-kvadrat* statistike, Arango predlaže pravljenje tablice 5 nepredviđenih slučaja (ili samo slučaja) *2 * c* gdje *c* predstavlja broj grupa (podsjetnik: *c1* je iz primjera ranije: „*Upisao/la =Da ∧ Godine 20-25*“) (Arango, bez dat.).

Tablica 5. Izgenerirani podaci potrebni za izračun

| | ***Računalne znanosti*** | ***Biologija*** | ***Inženjerstvo*** |
|---|---|---|---|
| ***c1*** | 11 | 15 | 18 |
| ***¬c1*** | 33 | 11 | 50 |

(Izvor: Arango, bez dat.)

*Hi-kvadrat* izračunava se formulom: $x^2 = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

*Hi-kvadrat* koristi se kod provjere da li dobiveni rezultat odstupa od očekivanog na primjer kod izrade simulacija. Postoji test pravih podataka koji su negdje prikupljeni, recimo u bazi podataka i simulirane podatke. *O* iz formule su simulirani podaci, a *E* iz formule , prikupljeni su podaci. Ovdje se računa samo vrijednost za *E* čija je formula u nastavku. Ako je vrijednost za *E* prihvatljiva na primjer ispod pet posto, onda govorimo o dobroj simulaciji. Dakle, *E* u nastavku zbroj je plavo ispisanih elemenata pomnoženih sa zbrojem zeleno ispisanih elemenata kroz šest jer je konačan broj promatranih slučajeva iz tablice šest.

$$E_{ij} = \frac{\sum_{i=1}^{2} O_{ij} \sum_{j=1}^{c} O_{ij}}{N}$$

Rezultati podrške u stupcu *Računalne znanosti* i rezultati podrške u retku *c1*, potrebni su za izračun očekivanih vrijednosti za redak jedan i stupac jedan, odnosno $E_{11}$.

U *hi-kvadrat* testovima, alfa (*α*) obično je 0.05. Alfa je maksimalna mogućnost za lažno odbacivanje nulte hipoteze (pogreška tipa I). Pogreška tipa I definirana je odbacivanjem istinite nulte hipoteze. Konkretno, od 1000 testova, u prosjeku će biti 50 pogrešaka tipa I.

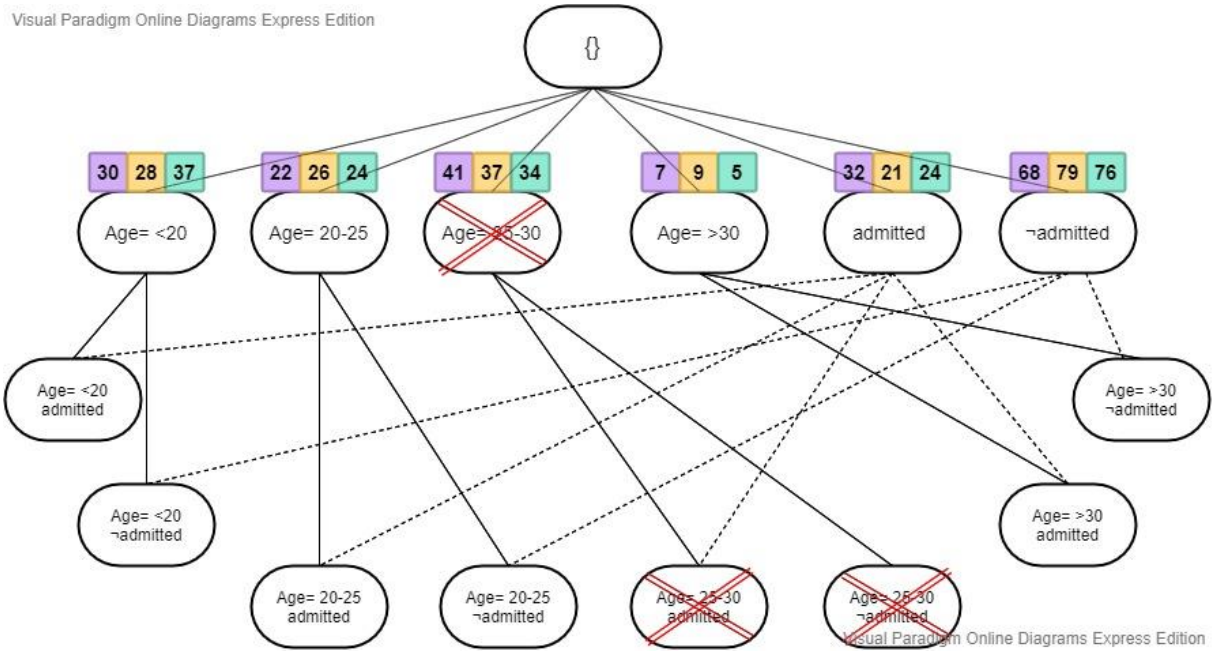
Postupno smanjivanje vrijednosti za alfu na svakoj razini, rješenje je koje predlaže autor Arango (bez dat.) u svom primjeru.

Slijedeće što algoritam radi jest da odlučuje da li koji čvor treba biti obrezan. Algoritam obrezuje, kako to Bay i Pazzani (bez dat.) navode, kada odluči da će traženje samo dovesti do nezanimljivih razlikovnih skupova, a potrebni su *zanimljivi* i *značajni*.

Postoje barem 3 strategije obrezivanja koje prezentira (Arango, bez dat; Bay i Pazzani, 2001):

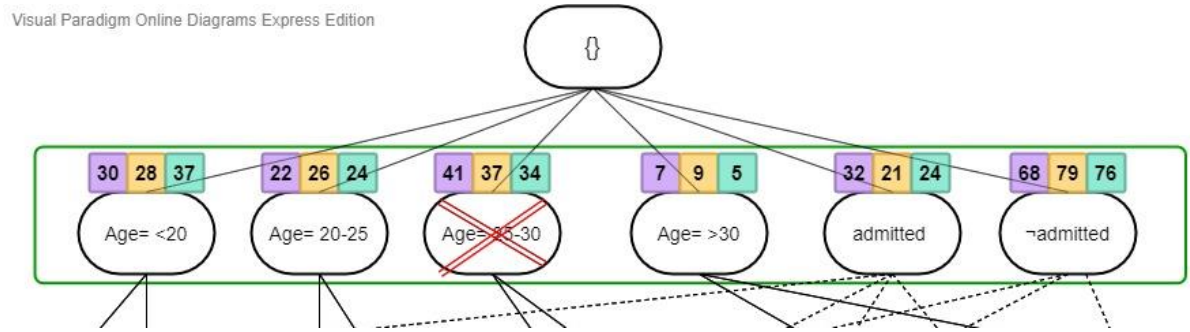
1. *Minimalna veličina odstupanja*- ako je podrška za čvor manja od parametra *mindev* za sve grupe
2. *Očekivana frekvencija stanica*- ako je očekivanje za čvor premalo, *hi-kvadrat* nije valjan, isto tako će biti nevažući za djecu
3. *Hi-kvadrat granice*- moguće je izračunati gornju granicu za *hi-kvadrat* statistiku za svu djecu nekog čvora. Ako nije dovoljno visoko da premaši alfa test za tu razinu, čvor se može obrezati.

Na slici 3. grafički je prikazano takozvano „obrezivanje“ koje je spomenuto ranije u tekstu.



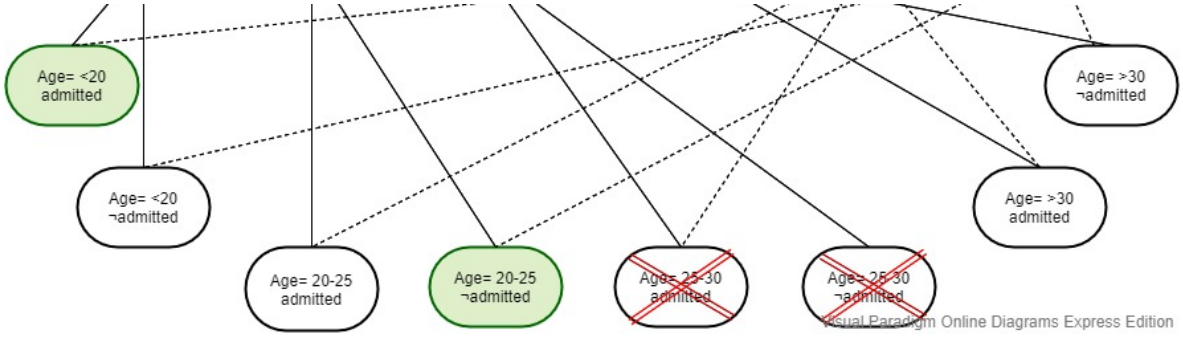

Slika 3. Prikaz „obrezivanje“ (Prema: Arango, bez dat.)

Algoritam potom prikazuje sva odstupanja prvog reda što je vidljivo na slici 4.




Slika 4. Prikaz odstupanja prvog reda (Prema: Arango, bez dat.)

STUCCO algoritam nakon toga prikazuje sva ostala odstupanja samo ako su ona iznenađujuća (slika 5). Razlikovni skup smatra se *iznenađujuć* ako je podrška skupa različita od onoga što se očekuje.




Slika 5. Ostala odstupanja (Prema: Arango, bez dat.)

Razni autori, poput Arange, Bayja i Pazzanija, zaključuju da STUCCO algoritam ispituje statističke hipoteze kako bi pronašao značajne razlike među grupama. Isto tako, pruža kontrolu nad lažnim-pozitivnim rezultatima, implementira različite tehnike obrezivanja odnosno postupka koji eliminira nezanimljive razlike te sažima rezultate (Arango, bez dat.)

## 2.2 Magnum Opus algoritam

Magnum Opus ili u prijevodu 'remek' djelo, dostojan je naziv provedbe algoritma Opus. Naziv OPUS otkriva mogućnost algoritma, a to je 'optimalno obrezivanje za pretraživanje nesortiranog'. Bazu čine grananja i pretraživačke veze koje omogućuju dopušteno učinkovito pretraživanje po prostoru u kojem unos redoslijeda operatora nije važan. Učinkovita pretraživačka sposobnost algoritma, demonstrirana je uzevši u obzir da strojno učenje (eng. *machine learning*) ima jako veliki pretraživački prostor. Korištenje takve dopustive pretrage, od značaja je za zajednicu strojnog učenja, što znači da se pristranost u učenju može primijeniti na složene zadatke te može biti konkretizirano i manipulirano. OPUS isto tako ima potencijala u aplikacijama koje su povezane s drugim područjima umjetne inteligencije (Webb, 1995).

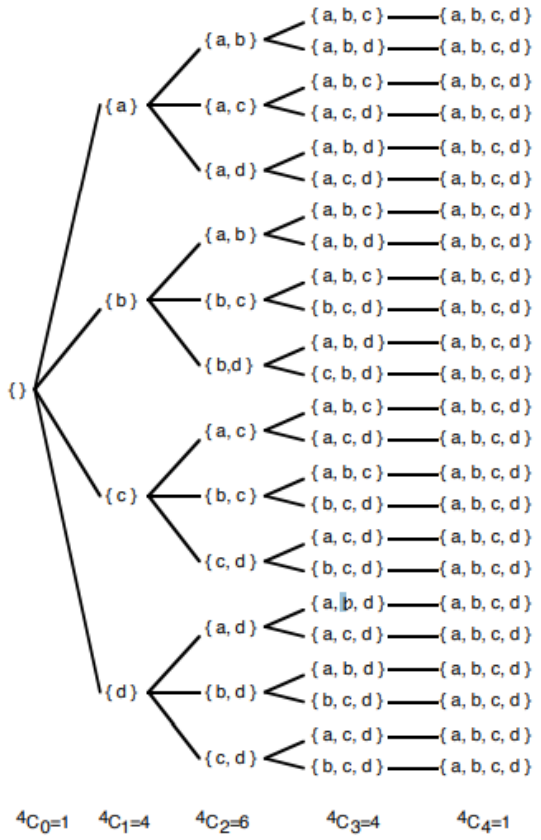
Mnogi problemi vezani uz umjetnu inteligenciju zahtijevaju pretraživanje. Pretraživački algoritmi donose učinkovito dopustivo pretraživanje pretraživačkog prostora pomoću što bolje pretraživačke učinkovitosti. Takva pretraživačka učinkovitost postignuta je korištenjem grananja i veznih tehnika koje vladaju specifičnim pravilima obrezivanja kako bi postigli usko usredotočen prelazak po pretraživačkom prostoru (Webb, 1995).

Što se tiče unosa redoslijeda operatora, OPUS algoritam je specifičan jer prilikom njegovog korištenja, unos redoslijeda operatora nije važan, kako je spomenuto u uvodu. Na primjer, kada se pretražuje prostor logičkih izraza, učinak spajanja izraza *A* s izrazom *B*, a zatim spajanja rezultata s izrazom *C*, identičan je rezultatu dobivenom spajanjem *A* s *C* iza kojeg slijede *B*. Obje sekvence operacija rezultiraju izrazima s jednakim značenjem. Općenito,

prostor za pretraživanje je nesortiran ako za bilo koji slijed *O* operatora i bilo koje stanje *S*, sva stanja koja se mogu koristiti od *S* zamjenom mjesta, identična su (Webb, 1995).

OPUS algoritmi prolaze prostor pretraživanja pomoću stabla pretraživanja (jednostavno nesortirano pretraživačko stablo na slici 6). Korijen pretraživačkog stabla je stablovo početno stanje. Grane označavaju primjenu pretraživačkih operatora i čvorova koji vode do označavanja stanja koje rezultira primjenom tih operatora. Za optimalno pretaživanje, cilj je optimalno rješenje. Za zadovoljavajuću pretragu, cilj je prihvatljivo rješenje. Moguće je da pretraživački prostor uključuje više različitih ciljeva.

Spomenuti algoritmi znaju iskoristiti svojstva nesortiranog pretraživačkog prostora kako bi optimizirali efekt bilo kakvog obrezivanja pretraživačkog stabla koje bi se moglo pojaviti. OPUS algoritmi ne sadržavaju pravila za obrezivanje. Radije ih koriste kao input i nastoje optimizirati učinak svakog prezivanja koje rezultira prilikom aplikacije svakog pravila obrezivanja. Algoritmi su dizajnirani s prihvatljivim pravilima obrezivanja. Kada se jedino koriste s prihvatljivim pravilima obrezivanja, algoritmi su prihvatljivi što zapravo znači da će garantirano pronaći neki cilj ako isti postoji u pretraživačkom prostoru. Iako, algoritmi se mogu koristiti s neprihvatljivim obrezivačkim otkrićima da bi se zadržala učinkovitost neprihvatljivih pretraživanja (Webb, 1995).




Slika 6. Jednostavno nesortirano pretraživačko stablo (Izvor: Webb, 1995, str. 434)

Webb (1995) i Oreški (2014) u svojim radovima navode da u je u biti Magnum Opus algoritma, „…upotreba *k*-optimalnih tehnika za otkrivanje povezanosti.“ (Oreški, 2014). Te tehnike omogućuju korisniku da izabere što je njemu interesantna povezanost i koliko veza odnosno *k-ova* želi pronaći. Zadaća algoritma pronalazak je *k* najboljih veza koje su prethodno definirane zahtjevima koje je postavio korisnik. Oreški (2014) piše da korisnik može odabrati jedan od kriterija mjerenja najboljih veza, a to su: moć, podrška, učinak, zanimanje i zastupljenost. Opis spomenutih mjera slijedi u nastavku.

Neka je *D* dani skup podataka, *A*= *LSP*- lijeva strana pravila (eng. *left hand side- LHS*), a *B*= *DSP*- desna strana pravila (eng. *Right Hand Side- RHS*). (Oreški, 2014).

Mjere za pravila, uzevši u obzir gore napisano, izgledaju ovako:

- ***zastupljenost*** (eng. *coverage*)- omogućuje uvid u broj slučaja kod kojih se stavka nalazi s LSP

$$zastupljenost(A \rightarrow B, D) = zastupljenost(A, D)$$

- **podrška** (eng. *support*)- prikazuje broj slučajeva u kojima se odnos između lijeve i desne strane ponavlja

$$podrška\ (A \rightarrow B, D) = zastupljenost\ (A \cup B, D)$$

- **moć** (eng. *strength*)- moć se dobije tako što se podijeli iznos podrške s iznosom pokrivenosti. Dobiveni broj pokazuje procjenu vjerojatnosti u kojoj će se stavka u DSP, prikazati onda kada će se prikazati i LSP

$$moć(A \rightarrow B, D) = \frac{podrška(A \rightarrow B)}{zastupljenost(A) * zastupljenost(B)}$$

- **zanimanje** (eng. *Lift*)- zanimanje se dobije tako što se iznos podrške podijeli s iznosom *podrške* koja bi postojala kad ne bi postojala veza između LSP i DSP; na jaču povezanost nam ukazuje viša vrijednost, a na slabiju povezanosz upućuje niža vrijednost

$$zanimanje\ (A \rightarrow B, D) = \frac{podrška(A \rightarrow B)}{zastupljenost(A) * zastupljenost(B)}$$

- **učinak** (eng. *leverage*)- brojka dobivena razlikom iznosa podrške s iznosom da ne postoji veza između stavki lijeve i desne strane

$$učinak\ (A \rightarrow B, D) = zastupljenost\ (A\ \rightarrow B, D) * (moć\ (A\ \rightarrow B, D) -\ zastupljenost\ (B, D))$$

**P** se naziva, prema Oreški (2014), rezultatom statističke procjene značajnosti pravila. Ako je niža vrijednost, manja je vjerojatnost da je pravilo neispravno. To može biti slučaj ako su lijeva i desna strana nekonektirane ili zato što jedna/više stavki na lijevoj strani ne doprinosi povezivanju sa stavkom na desnoj strani.

Oreški (2014) primjećuje da „…ako je *p<0.05*…" ono je statistički značajno. U skladu s time „…ako je *p>0.05*, pravilo nije statistički značajno te postoji vjerojatnost da je rezultat slučajnosti, a ne stvarnog stanja." (Oreški, 2014).

Vrijedi spomenuti test pomoću kojeg Magnum Opus pronalazi razlikovne skupove. *Binomni test* koristi se pri provjeri podataka koji imaju samo dva moguća krajnja rezultata. Za slučajeve koji imaju samo dva moguća krajnja rezultata, svi podaci moraju biti u jednoj od diskretnih skupina. Primjeri skupina su: neispravan-ispravan, žensko-muško, događaj X-događaj Y. Poznavanjem vjerojatnosti jedne skupine *P*, poznajemo i vjerojatnost druge skupine *1-P*. Pretpostavljanjem i poznavanjem vjerojatnosti populacije P, može se testirati da li uzorak dolazi iz te populacije (Dujman, 2017).

Zaključuje se da binomni test ukazuje da li vjerojatnosti ili frekvencije promatranog uzorka, mogu biti iz raspodjele s određenom vrijednošću *P*.

Primjerice, kolika je vjerojatnost dobivanja dvije šestice prilikom bacanje kocke 5 puta? Broj mjerenja u ovom primjeru je *N (N=5)*, broj šestica je oznaka *x*, odnosno *x=2*, vjerojatnost da se baci baš šestica je 1/6, dakle *p=1/6* (jer je na kocki šest različitih vrsta brojka).
Vjerojatnost dobivanja dvije šestice jest prikazano je jednadžbama:

$$\mathrm{P_x} = \binom{\mathrm{n}}{\mathrm{x}}\mathrm{p^x q^{n-x}}$$

$$\mathrm{P_2} = \binom{5}{2}\mathrm{p^2 q^{5-2}} = \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = 0{,}16$$

Dujman (2017) pak navodi ako se pokaže zanimanje za vjerojatnost da li je *x* veći ili manji od specifične vrijednosti, sumira se vjerojatnost događanja *x-a*: Šanse da se dobiju dvije ili manje šestica *p(x ≤ 2)* vjerojatnost je da se dobije ni jedna šestica *p(x=0)*, jedna šestica *p(x=1)* i dvije šestice *p(x=2)*. Potom se sumiraju vjerojatnosti: *p(x ≤ 2)= p(x=0) + p(x=1) + p(x=2).*

$$P_{x=} \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x}$$

$$p(x < 2) = p(0) + p(1) + p(2) = 0.40 + 0.40 + 0.16 = 0.96$$

Vjerojatnost dobivanja dvije ili manje šestica prilikom bacanja kocke 5 puta je 96%.

# 3. Usporedba algoritama

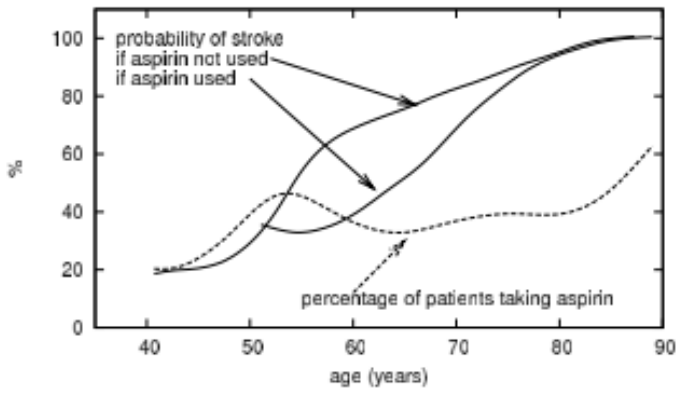
U poglavlju „Usporedba algoritama“ rezimirane su glavne razlike između spomenutih tehnika otkrivanja kontrasta: STUCCO i Magnum Opus algoritma.

Osnovna primjetna razlika između tehnika, očito je način na koji algoritmi primjenjuju sortiranje za pronalazak razlikovnih skupova. STUCCO primjenjuje hi-kvadrat test koji je objašnjen ranije, a Magnum Opus, pošto ima samo dva moguća ishoda, grupe, koristi binomni test. Možda bi bilo sigurnije koristiti STUCCO algoritam kada je potrebno otkriti razlike odnosno kontraste među grupama pošto je delikatniji na uži raspon drastičnih oblika kontrasta što nužno ne znači da je će nam STUCCO algoritam uvijek dati točnije odgovore od Magnum Opus algoritma. Način sortiranja koji koristi Magnum Opus, blaži je od STUCCO-vog jer STUCCO provodi usporedbu između grupa, a Magnum Opus izvodi usporedbu u unutrašnjosti grupa. Nadalje, STUCCO posjeduje ograničenje kod minimalne veličine kontrasta između grupa i upotrebljuje ispravke (eng. *Bonferonni correction*) nad višestrukim usporedbama i provedbama pronalaska minimalnog kontrasta (u Magnum Opus algoritmu ne postoji takav ispravak).

# 4. Primjena algoritama

Poglavlje primjene algoritama, opisuje autorici završnog rada, dosadašnje najzanimljivije, uspješno primjenjivanje tehnike otkrivanja kontrasta. Radi se o djelu originalnog naziva autora Nazari i suradnika: „Analyzing *Relationships Between Aircraft Accidents and Incidents*“ koji analizira veze između avionskih nesreća i incidenata pomoću STUCCO algoritma. (Nazari et. al, 2008).

Vrijedi spomenuti također zanimljivo i korisno istraživanje Kralja i suradnika: „*Contrast Set Mining Through Subgroup Discovery App*“, koje govori o pacijentima s moždanim udarom, razlici istih u odnosu na pacijente koji imaju drugačije neurološke poremećaje, a iste simptome kao i pacijenti koji su doživljeli moždani udar. Jedna analiza vjerojatnosti od moždanog udara, uvjetovana vjerojatnošću pacijenata koji su imali moždani udar, pokazala je sjajnu motivaciju pacijenata iznad pedesetidvije godine, da počinju preventivno uzimati terapiju aspirinom jer algoritam eksplicitno prepoznaje važnost uzimanja terapije aspirinom nakon 52. godine života. Gornja krivulja prikazuje vjerojatnost dobivanja moždanog udara ako se aspirin ne uzima, donja krivulja ako se aspirin uzima, a iscrtkana krivulja postotak pacijenata koji uzimaju aspirin. (Kralj et. Al., 2007).




Slika 7. Vjerojatnost moždanog udara (Izvor: Kralj et. al., 2014, str. 9)

U nastavku poglavlja slijedi sažetak izabrane spomenute primjene.

Sustavi zračnog transporta napravljeni su tako da se mogućnost nesreće smanji na minimum. Da bi se to ostvarilo potrebno je razumijeti tj. istražiti faktore koji utječu i pridonose mogućnosti nesreće, međutim, mali broj nesreća otežava pronalaženje tj. prepoznavanje ponavljajućih uzoraka tih faktora. Upravo je to cilj ovog istraživanja, korištenje tehnika

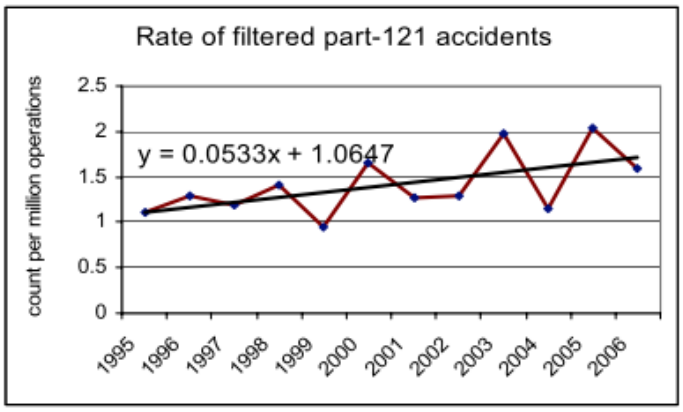
rudarenja podataka za pronalazak, prepoznavanje i povezivanje podataka i faktora iz zrakoplovnih incidenata i nesreća.

Razina sigurnosti mjeri se brojem nesreća i incidenata.

Zrakoplovna nesreća događaj je koji je povezan s upravljanjem zrakoplova, a koji kao posljedicu ima ozljeđene ili mrtve osobe i/ili u kojem avion pretrpi priličnu štetu.

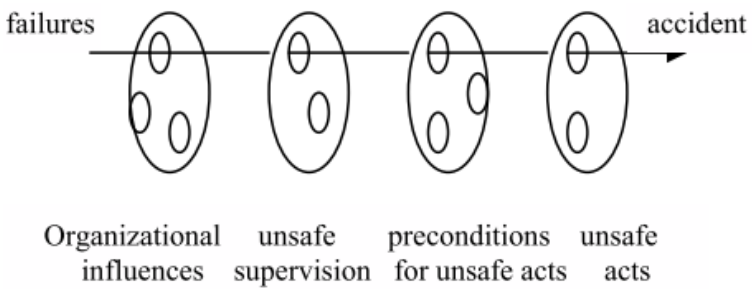
Zrakoplovni incident događaj je koji nije zrakoplovna nesreća, ali je potencijalno opasan i koji je uz pojavu jednog ili više faktora mogao rezultirati ozlijedom, smrtnim slučajem ili priličnom štetom na zrakoplovu.

Zrakoplovni promet je sve sigurniji, no ima još potrebe za većom sigurnošću, kao što je očito iz slike 8. na kojem je prikazana povezanost između broja letova i broja nesreća.



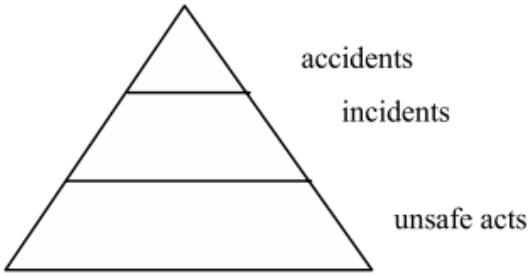

Slika 8. Povezanost između broja letova i broja nesreća (Izvor: Nazeri et. al, 2008, str. 185)

Dva glavna pristupa u istraživanjima sigurnosti su model „*švicarskog sira*“ i „*Heinrichova piramida*“. Model švicarskog sira (slika 9), nesreću opisuje kao prolazak kroz rupe u švicarskom siru s time da švicarski sirevi predstavljaju sisteme koji su tu da spriječe štetu. Većina događaja se zaustavi na vrijeme, ali ima i događaja gdje se rupe na sirevima poklope i događaj rezultira nesrećom.




Slika 9. Model „švicarskog sira“ (Izvor: Nazeri et. al, 2008, str. 185)

Heinrichova piramida (slika 10), opisuje nesreće u smislu frekvencija. Na dnu piramide opisani su postupci koji su najčešći, iznad njih su incidenti koji su malo rjeđi, a na vrhu su nesreće koje su najrjeđe. Po ovom modelu dobiveno je da za svaku veću nesreću ima 3-5 nefatalnih nesreća, 10-15 incidenata i stotine neprijavljenih događaja.




Slika 10. Prikaz modela Heinrichove piramide (Izvor: Nazeri et. al, 2008, str. 185)

Ovo istraživanje, kao i Heinrichova piramida, traži povezanost između nesreća i incidenata, ali ne na kvantitativnoj razini već kroz faktore koji su prisutni u obje grupe događaja.

Korišteni su podaci o nesrećama i incidentima na komercijalnim letovima u periodu od 1995. do 2004. godine. Podaci su dobiveni od američkog nacionalnog odbora za sigurnost transporta (eng. *National Transportation Safety Board* – NTSB) dok su podaci o incidentima dobiveni od agencija *FAA/AIDS, NASA/ASRS, FAA/OED i FAA/SDRS*. Svaki izvještaj sadrži opis događaja i faktora koji su uzrokovali ili doprinjeli, a koji identificiraju ili osoba koja je pisala izvještaj ili stručno lice koje pregledava izvještaj.

Cilj analize bila je identifikacija uzoraka faktora nesreće koji su bili povezani s rutinskim upravljanjem zrakoplova. Zbog toga se incidenti i nesreće kao posljedice sljedećih uzroka ne uzimaju u obzir:

- medicinski ili alkoholni slučajevi, npr. pilot je pio pod utjecajem alkohola
- terorizam i slučajevi narušene sigurnosti, npr. prijetnja bombom
- slučajevi povezani sa putnicima ili osobljem aviona, npr. putnik je pretrpio ozljede uslijed proljevanja tople kave
- udar životinja/ptica
- događaji koji su se desili kada zrakoplov nije bio u upotrebi (parkiran, provjere pred let)

Sve nesreće civilnih zrakoplova u SAD istražuju se od strane NTSB i pohranjuju se u njihovu bazu. Kako je NTSB neovisna organizacija, svi podaci vezani za nesreće mogu se smatrati potpunim i nepristranim. S druge strane, incidenti se rjeđe službeno prijavljuju i istražuju te postoji mogućnost pristranosti. Da bi se smanjio utjecaj pristranosti u izvještajima,

u ovom istraživanju uzimani su u obzir i primarni faktori i faktori koji su doprinjeli nekom događaju.

Prvo je stvorena zajednička taksonomija za baze nesreća i incidenata kako bi se prepoznali zajednički faktori iz obje baze. Potom je svaki izvještaj pretvoren u vektor matricu s odgovarajućim poljima koja su popunjavana odgovarajućim vrijednostima svakog izvještaja. Na vektore (matrice) nesreća i incidenata je primjenjen STUCCO algoritam koji prepoznaje uzorke faktora koji pokazuju jaku povezanost s nesrećama ili incidentima. Rezultati su rangirani pomoću *faktora Odnos-podrška* (eng. *Factor Support Ratio*), ovaj pristup pobliže je objašnjen u daljnjem tekstu. Rezultati analiza različitih baza izvještaja, uspoređivani su kako bi se potvrdila efektivnost pristupa.

Proces stvaranja zajedničke taksonomije bazira se na analizi podataka. Nakon analizire struktura svih baza podataka i vrijednosti za svako polje, razvila se hijerarhija faktora i podfaktora koji se pojavljuju u raznim bazama podataka. Faktori se mogu razvrstati u osam glavnih kategorija koje onda sadržavaju podfaktore. Kategorija Drugi (eng. *Other*), sadrži sve podfaktore koji nisu dovoljno veliki da bi imali svoju kategoriju. Ovaj raspored je prikazan u tablici 6.

Tablica 6. Primjeri faktora i podfaktora

| Faktor | Primjer za pod-faktor |
|---|---|
| **Zrakoplovstvo** | Motor, sustav za kontrolu leta, oprema za slijetanje |
| **Aerodrom** | Neočišćeni snijeg s piste, slabo osvjetljenje, zbunjujuće oznake |
| **Kontrola leta** | Komunikacija s pilotom, nošenje s procedurom |
| **Kompanija** | Procedure, rukovođenje, obuka |
| **Održavanje** | Usklađenost, inspekcija |
| **Pilot** | Vizualno snalaženje, visinsko odstupanje, odluka/procjena |
| **Vrijeme** | Vjetar, grmljavina, led |
| **Drugo** | Faktori koji nisu u drugim kategorijama; vidljivost |

(Prema: Nazeri et. al., 2008, str. 187)

Bilo je potrebno izvršiti normalizaciju vrijednosti kako bi sve baze podataka koristile istu riječ tj. frazu za isti faktor tj. stanje. Na primjer, za postupak gdje je pilot trebao izvesti manevar kako bi izbjegao vozilo ili objekt na pisti, je u jednoj bazi korišten termin susret na tlu (eng. *ground encounter*) dok je u drugoj korišten izbjegavanje objekta (eng. *object avoidance*)
Izvještaji su potom pretvarani u vektor (matricu) koja sadržava polja koja pokazuju prisutnost tj. odsutnost svakog faktora i podfaktora u nesreći ili incidentu. Ovi vektori se potom koriste u analizi.

Kako je cilj istraživanja identifikacija i prepoznavanje faktora i kombinacija faktora koji prethode nesrećama, bilo je potrebna analitička tehnika koja bi uzimala u obzir oba skupa

podataka (nesreća i incidenata) i koja bi mogla odrediti koji faktori imaju veću vjerojatnost da rezultiraju nesrećom. STUCCO algoritam primijenjen je na vektore matrice nesreća i incidenata kako bi ih analizirao pomoću kontrasta. Algoritam pronalazi čvorišta parova atribut-vrijednost koji su prilično različiti u različitim grupama. U ovom slučaju, postoje dvije grupe, nesreće i incidenti. Parovi atribut-vrijednost binarne su vrijednosti koje ukazuju na prisutnost ili odsutnost faktora u nekom događaju. Analizirana je učestalost faktora i njihove „djece" (kombinacija faktora) u svakoj grupi. Za svaki skup faktora, devijacija tj. odstupanje dobiveno je razlikom vrijednosti učestalosti nekog faktora u nesrećama i incidentima. Prvi korak filtriranje je skup faktora koji ne prelaze minimum vrijednosti devijacije u ovom slučaju je određeno da bude 1%. Sljedeći korak provođenje je već spomenutog *hi-kvadrat* testa, kako bi se provjerio statistički utjecaj skupa faktora kroz dvije glavne grupe. Korištena je tablica slučaja pod brojem 7 koja se koristi za hi-kvadrat test značajnosti. Granica *p*-vrijednosti je određena da bude 0.05. Skup faktora (eng. *factor set*) koji imaju *p*-vrijednost veću od 0.05 se odbacuju. *P*-vrijednost koja je manja od 0.05 je jednaka 95% sigurnost i ona se prihvaća.

Tablica 7. Tablica slučaja

| | **Nesreće** | **Incidenti** |
|---|---|---|
| **skup faktora istinitosti** | nesreće koje sadrže skup faktora | incidenti koji sadrže skup faktora |
| **skup faktora neistinitosti** | nesreće koje ne sadrže skup faktora | Incidenti koji ne sadrže skup faktora |

(Prema: Nazeri et. al., 2008, str. 187 )

Kada su prepoznati utjecajni skupovi faktora uz pomoć algoritma, rangirani su pomoću faktora Odnos-podrška. Kako se vidi iz jednadžbe niže, izračunava se za svaki komplet faktora kao odnos podrške skupa faktora u skupu podataka nesreća i podrške skupa faktora u skupu podataka incidenata.

$$Odnos\ podrške = \frac{podrška_{\ nesreća}}{podrška_{incident}}$$

*Odnos podrške* je vjerojatnost da skup faktora bude prisutan u nesreći podjeljena s vjerojatnoćom da bude pristutan u incidentu. Rezultat ovog postupka, drugačiji je od devijacije (razlika između podrške skupa faktora za nesreću i incidente). Kako bi se odnos podrške bolje shvatio, može ga se primjeniti na skup faktora *A* i *B* u tablici 8.

Tablica 8. Podaci za shvaćanje odnosa podrške

| Skup faktora | Podrška nesreća | Podrška incidenta | Odstupanje | Odnos podrške |
|---|---|---|---|---|
| A | 60% | 50% | 10% | 1.2 |
| B | 11% | 1% | 10% | 11 |
| C | 60% | 10% | 50% | 6 |

(Prema: Nazeri et. al., 2008, str. 187)

Oba skupa faktora imaju odstupanje od podrške u nesrećama i incidentima koje iznosi 10%. Za slučaj skupa faktora *B*, podrška za nesreće je 11 puta veća nego za incidente. To se može protumačiti na sljedeći način- mogućnost prisutnosti skupa faktora *B* u nesreći je 11 puta veća nego mogućnost njene prisutnosti u incidentu. Ovo je uočljiviji raspored nego raspored skupa faktora *A* čiji je *Odnos podrške* 1.2. Mjera se može upotrijebiti pri uspoređivanju skupa faktora *A* i *B* i za njihov slučaj može se reći da skup faktora *B* ima veću vjerojatnost da bude prisutan tj. uključen u nesrećama nego skup faktora *A* (Nazeri et. al., 2008).

Da bi se shvatila važnost *Odnosa podrške* u rangiranju faktora, može se uzeti slučaj skupa faktora *A* i *C* u tablici iznad. Oba skupa faktora pojavljuju se u 60% nesreća, ali činjenica da se skup faktora *C* pojavljuje u 10% incidenata povisuje njegov *Odnos podrške* (u odnosu na skup faktora *A*). Može se protumačiti kao: skup faktora *C* kvalitativno je utjecajniji faktor nesreća nego skup faktora *A*. Kada se skup faktora *C* pojavi, onda je veća vjerojatnost (6 puta veća) da to bude nesreća nego incident, ali kada se skup faktora *A* pojavi, vjerojatnost nesreće u odnosu na incident je manji (1.2).

*Podrška faktora nesreći* (učestalost faktora u nesrećama) pokazuje koliko puta je faktor bio prisutan u nesrećama, ali ne pokazuje koliko često se faktor pojavio tj. desio (u nesrećama i incidentima). Slično tome, *podrška faktora* incidenta pokazuje koliko puta se faktor pojavio (bio prisutan) u incidentima, ali bez indikacije njegove uloge u incidentima (da li je bio utjecajan, u kojoj mjeri). U nekim slučajevima, faktori koji su često prisutni u incidentima rijetko su prisutni u nesrećama. Ovo ukazuje na to da taj faktor nije utjecajan faktor u nesrećama. Jedno od objašnjena može biti da je taj faktor zaustavljen kada se pojavio, a prije nego što je mogao dovesti do nesreće.

Nazeri et. al. izvršavaju zasebne analize na 4 para skupa podataka. Svaki par sastojao se od izvještaja nesreće i odgovarajućih izvještaja iz jedne od 4 baze podataka incidenata. Rezultati analiza uspoređeni su na kraju. Dalje će se predstaviti rezultati istraživanja koji su potvrđeni višestrukom analizom parova baza podataka incident/nesreća.

Vjerojatnost da faktori dovedu do nesreće (za razliku od incidenta) veća je ako su iskombinirani. Rangiranje rezultata pomoću *faktora Odnosa podrške*, pokazuje da vjerojatnost prisutnosti faktora u nesreći raste s pojavom drugih faktora. Na primjer, *Odnos podrške* za kombinaciju faktora pilot+aerodrom je 7.2 u odnosu na *Odnos podrške* za faktor pilot koji je

sam po sebi 3.8. Ovo ukazuje da kombinacija faktora pilot i aerodrom ima 1.8 puta veću vjerojatnost da rezultira nesrećom nego samo faktor pilot.

*Faktori kompanije* obuhvaćaju greške koje je počinilo osoblje kompanije tj. zrakoplovne linije, nedovoljne ili ne postojanje procedura kompanije pri obavljanju nekog zadatka, nedostatak upravljanja tj. menadžmenta od strane menadžmenta kompanije. Između osam kategorija faktora u podacima, faktori kompanije bili su najviše rangirani pri rangiranju faktora pomoću odnosa podrške.

Sljedeća najviša rangirana kategorija nakon kategorije faktori kompanije je bila kategorija *kontrole leta* (eng. *air traffic control*). Najutjecajniji podfaktor u ovoj kategoriji bila je komunikacija kontrole leta koja se odnosi na izdavanje savjeta od strane kontrolera, davanje informacije o vremenu pilotu od strane kontrolera i provjera točne potvrde informacije pilota kontroleru (eng. *readback*).

*Faktori pilota* češći su od drugih faktora u nesrećama, ali se češće pojavljuju i u incidentima i zbog toga je njihov odnos podrške manji te se ih rangira nakon faktora kompanije i faktora kontrole leta. Najutjecaniji podfaktor vizualno je promatranje.

*Zrakoplovni faktori* odnose se na mehaničke probleme zrakoplova, njegovih komponenti ili sustavi. Primjeri su problemi s opremom za slijetanje, sustavima za kontrolu leta i krilima. Bez prisutnosti drugih faktora, zrakoplovni faktori klasificiraju se kao faktori incidenata što znači da je veća vjerojatnost da uzrokuju incident nego nesreću (ako se pojavljuju sami). Kada su zrakoplovni faktori u kombinaciji sa drugim faktorima (npr. teški vremenski uvjeti ili greške pilota), ta kombinacija postaje faktor nesreće (Nazeri et. al, 2008).

Analiza podataka desetogodišnjeg perioda (1995-2004) pokazala je da *faktori pilota* i *faktor zrakoplova* opadaju i da *faktori kontrole leta* rastu. Izvještaji ukazuju na to da je rast *faktora kontrole leta* uvjetovan raznim uvjetima koji se nazivaju *faktorima kompleksnosti* tj. složenosti. Dostupni podaci sadržavali su jedanaest ovih *faktora složenosti*: dizajn zračnog prostora, hitni događaj, iskustvo kontrolera, kontrola protoka, broj zrakoplova, uvjeti piste, konfiguracija tj. raspored piste, teren, posebni događaji, vrijeme i drugi. Podaci desetogodišnjeg perioda pokazuju da su; broj zrakoplova, dizajn zračnog prostora, konfiguracija piste i iskustvo kontrolera, najčešći faktori kompleksnosti koji utječu na faktor kontrole leta.

Nazeri et al. (2008) govore da se može očekivati da uvjet složenosti broja zrakoplova raste s obzirom da je predviđen i daljnji rast zračnog transporta. Ovaj rast će imati indirektni utjecaj i na konfiguraciju piste i na dizajn zračnog prostora. Zbog povećanog protoka, zračne luke trebati će koristiti konfiguracije koje omogućavaju veći broj dolazaka i odlazaka. Poteškoće vezane uz dizajn zračnog prostora, na primjer, ograničen prostor za promjenu visine pri prelasku zrakoplova iz jednog zračnog prostora u drugi, biti će naglašenije povećanjem broja zrakoplova. Faktor *iskustvo kontrolera* također će biti pod utjecajem rasta

broja zrakoplova jer će biti potrebe za većim brojem kontrolera koji će se moći nositi s povećanim volumenom posla. Uz to, broj kontrolera koji su otišli u mirovinu u zadnjih par godina, bio je veći od predviđenog i previđa se umirovljivanje velikog broja kontrolera u sljedećih par godina Federalna Agencija za Avijaciju (eng. *FAA*) planira zaposliti 1000 kontrolera svake godine tijekom sljedećih 10 godina (Nazeri et. al, 2008).

Uzimajući u obzir faktora nesreća koji su istaknuti u ovom istraživanju, projicirani rast zrakoplovnog transporta i posljedice tog rasta kao i njihov utjecaj na faktore nesreća očekuje se da će rasti i učestalost nesreća ukoliko se ne uvedu promjene postojećih uvjeta (Nazeri et. al, 2008).

# 5. Vlastiti primjer primjene STUCCO algoritma

U poglavlju „Vlastiti primjer primjene STUCCO algoritma“, slijedi konkretna primjena algoritma STUCCO.

Postoji skup podataka koji ima neke atribute. Uzima se proizvoljan broj atributa (dva ili više) i provjerava se da li bi oni mogli biti kontrastni (razlikovni) set. Za te odabrane atribute računa se podrška. Nakon izračunavanja podrške za sve grupe, podrške se uspoređuju kako bi se vidjelo da li su značajne ili ne. Značajnost je vidljiva kroz hi-kvadrat. Hi-kvadrat računa se tako što se računa razlika između neke dvije vrijednosti, rezultat se kvadrira, zbraja i dijeli s brojem koji predstavlja ukupan broj vrijednosti u skupu.

Uzet je za primjer (vidi Excel list broj 3 u prilogu), skup podataka koji sadrži vrstu životinje, spol, godine_cijepljenja (ako vrijednost godine_cijepljenja nije unesena, za potrebe primjera, uzima se da životinja nije cijepljena i to predstavlja broj -1), gdje je bio ugriz životinje pa se grupira po tipu životinje i promatra se koliko grizu cijepljene životinje, a koliko ne cijepljene.

Još se smanjuje skup podataka pa se uspoređuju podaci za pasmine; labrador (LR), njemački ovčar (GS), chihuahue (CH), bokser (BO). Za prazne vrijednosti spola i mjesta ugriza, dodana je nepoznanica (eng. *Unknown*) u Excel listu broj 3.

Konkretno u primjeru, ***c je: Spol** = **muški** ∧ **godine_cijepljenja** = **-1*** (muški psi koji nisu cijepljeni).

*Podrška* za pojedine grupe izračunava se ovako:

$$\sup(c \mid BO) = \frac{38}{181} = 0{,}20994 \;\; \check{s}to\ je\ 20{,}99\ \%$$

$$\sup(c \mid CH) = \frac{47}{165} = 0{,}28485\ \check{s}to\ je\ 28{,}948\ \%$$

$$\sup(c \mid GS) = \frac{66}{327} = 0{,}201835 \;\; \check{s}to\ je\ 20{,}18\ \%$$

$$\sup(c \mid LR) = \frac{50}{253} = 0{,}197628 \;\; \check{s}to\ je\ 19{,}76\ \% \;\;,$$

gdje je kod boksera, 38 broj muških jedinki koji nisu cijepljeni, a 181 su svi bokseri, dakle, muški, ženski, cijepljeni, ne cijepljeni itd.) Isto vrijedi i za ostale grupe.

Da bi se izračunalo da li je skup (set) *Velik*, uzima se *max* i *min* vrijednost ***podrške***, odnosno razlika između postotka: 28,948 % – 19,76 % = 8,72 %. Set je *Velik* ako je razlika između navedenih postotaka veća od 5% (parametar *mindev*). Zaključuje se da je set *Velik*.

Za računanje značajnosti, potrebne su:

- *stvarne vrijednosti,*
- *očekivane vrijednosti,*
- *razlika stvarne i očekivane vrijednosti,*
- *kvadrat razlike*
- *te kvadrat razlike/očekivana vrijednosti* i
- *hi kvadrat* na kraju.

*Stvarna vrijednost* za *c* (tablica 9) za grupu boksera (eng. *BO*) je 38, kao što je gore zaključeno, a *stvaran negativna vrijednost za c* ($\neg c$) je broj kada od zbroja svih boksera (181), oduzmemo *stvarnu vrijednost* za *c* (38). Dobije se da je *stvarna negativna vrijednost c-a* 143 za grupu boksera. Isti postupak je i za ostale grupe.

Tablica 9. Stvarne vrijednosti za c i $\neg c$

| **stvarne:** | **BO** | **CH** | **GS** | **LR** |
|---|---|---|---|---|
| ***c*** | 38 | 47 | 66 | 50 |
| $\neg$ ***c*** | 143 | 118 | 261 | 203 |

(Izvor: vlastita izrada)

*N* je suma svih grupa, odnosno broj 926.

*Očekivana vrijednosti* za *c* (tablica 10), primjerice njemačkog ovčara (eng. *GS*) je suma njegovog *stvarnog c-a* i *stvarnog negativnog c-a* (66 i 261) pomnožena sa sumom svih *stvarnih c-eva* (38, 47, 66, 50) te podijeljena s brojem *N* (926). *Očekivana negativna vrijednost c-a* je suma njegovog *stvarnog c-a* i *stvarnog negativnog c-a* (66 i 261) pomnožena sa sumom svih *stvarnih negativnih c-eva* (143, 118, 261, 203) podijeljena s *N* = 926.

Tablica 10. Očekivane vrijednosti za za c i $\neg c$

| **očekivane:** | **BO** | **CH** | **GS** | **LR** |
|---|---|---|---|---|
| ***c*** | 39.28834 | 35.81533 | 70.97948 | 54.91685 |
| $\neg$ ***c*** | 141.7117 | 129.1847 | 256.0205 | 198.0832 |

(Izvor: vlastita izrada)

*Razlika za c* (tablica 11) je broj dobiven oduzimanjem *stvarne vrijednosti c-a od očekivane vrijednosti c-a. Razlika za negativan c je broj dobiven oduzimanjem stvarne negativne vrijednosti c-a* od *očekivane negativne vrijednosti c-a.*

Tablica 11. Razlika za c i $\neg c$

| razlika: | BO | CH | GS | LR |
|---|---|---|---|---|
| $c$ | -1.28834 | 11.18467 | -4.97948 | -4.91685 |
| $\neg c$ | 1.288337 | -11.1847 | 4.979482 | 4.916847 |

(Izvor: vlastita izrada)

*Kvadrat razlike za c* (tablica 12) broj je dobiven množenjem *razlike za c* s *razlikom sa c*, a *kvadrat razlike za negativan c* je *razlika za negativan c* na kvadrat.

Tablica 12. Kvadrat razlike c i $\neg c$

| kvadrat razlike: | BO | CH | GS | LR |
|---|---|---|---|---|
| $c$ | 1.659812 | 125.0967 | 24.79524 | 24.17538 |
| $\neg c$ | 1.659812 | 125.0967 | 24.79524 | 24.17538 |

(Izvor: vlastita izrada)

*Kvadrat razlike/očekivana vrijednost za c* (tablica 13) je *kvadrat razlike za c* podijeljen s *očekivanom vrijednosti za c*. *Kvadrat razlike za negativan c/očekivana negativna vrijednost c-a* je *kvadrat razlike za negativan c* podijeljen s *očekivanom negativnom vrijednošću c-a*.

Tablica 13. Kvadrat razlike/očekivana vrijednost za c i $\neg c$

| kvadrat razlike: | BO | CH | GS | LR |
|---|---|---|---|---|
| $c$ | 0.042247 | 3.492826 | 0.34933 | 0.440218 |
| $\neg c$ | 0.011713 | 0.968356 | 0.096849 | 0.122047 |

(Izvor: vlastita izrada)

Konačno se može izračunati *hi-kvadrat*, koji se dobijem sumom svih *kvadrata razlike/očekivanih vrijednosti za c* i *kvadrata razlike za negativan c/očekivana negativna vrijednost-c-a.* Iznos *hi-kvadrata* je 5,52358 što je veće od alfe ($\alpha$) pa se zaključuje da set nije značajan odnosno signifikantan.

Bio je ovo primjer primjene izračunavanja jednog seta pomoću STUCCO algoritma. Dalje se kreira stablo odlučivanja s kombinacijama dva ili više atributa koji se potom ispituju. Da bi se proces ubrzao, stablo se ponekad obrezuje, što možda nije uvijek najbolje rješenje jer se može desiti da se odreže kombinacija koja je kontrastni skup.

# 6. Zaključak

Kontrast u kontekstu ovog završnog rada označava promijenu, razliku ili usporedbu. Stoga kontrastno rudarenje podataka (eng. *contrast dana mining*) uključuje usporedbu jedne grupe objekata s drugom grupom objekata pomoću STUCCO i Magnum Opus tehnika otkrivanja kontrasta. Za STUCCO algoritam bitno je zapamtiti pojmove traženja *značajne* ili *iznenađujuća* razlike., *hi-kvadrat test* i *pronalazak razlika između grupa*. Magnum Opus *uspoređuje razlike u unutrašnjosti grupa*, koristi takozvane *Bonferonni korekcije* i primjenjuje *binomni test*. Dobivanjem razlika pomoću spomenutih tehnika, koristi se kako bi se uočile razlike između danih objekata pa se te informacije koriste kao dodatno, uvijek potrebno znanje. Razumijevanje po čemu se skupovi podataka razlikuju, nam može pomoći kako da određeno prikupljeno i dobiveno znanje koristimo u prikladnim situacijama poput medicinskih istraživanja značajnih za život ljudi ili pak samog spašavanja života.

Rudarenjem se poboljšava proces donošenja odluka na strateškoj i poslovnoj razini, pružajući uvid u „skrivene“ podatke business intelligence (eng. *BI*) metodologijom, otkrivaju se relacije, pravilnost, logičnost te općenito bilo kakve strukture među podacima. U poslovanju znatno pomaže bržem donošenju poslovnih odluka zbog kontaktiranja samo onih prospekata ili klijenata za koje postoji visoka vjerojatnost da će odgovoriti, odnosno vjerojatnost uspjeha. Kompleksne baze podataka, prepune su podacima, ali im nedostaju informacije koje su skrivene i spremljene u podacima. Rudarenje podataka pomaže otkriti znanje i važne informacije koje su utkane u podatke te uvelike pridonosi donošenju odluka, poslovanju i znanosti. Primjena rudarenja podataka prisutna je također u analizi košarice kupaca odnosno otkrivanje proizvoda čija je potražnja veća te se koristi i u genetičkim znanostima, bio-informatici, medicini i obrazovanju.

Područje rudarenja podataka već nudi i nudit će u budućnosti bezbroj mogućnosti pronalaska rješenja na postavljenje istraživačke pretpostavke te će svojim unapređivanjem i usavršavanje, tehnike otkrivanja kontrasta svakodnevno, još i kvalitetnije, poboljšavati živote svima nama.

# 7. Popis literature


Arango, C. (bez dat.). Mining for Contrasting Sets (STUCCO). Preuzeto 12.09.2020. s https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput695/F07/slides/contrastsets.pdf

Bay, S. D., Pazzani, M. J.(2001). Detecting Group Diferences: Mining Contrast Sets. Preuzeto 13.09.2020. s https://www.researchgate.net/publication/263112786_Detecting_Group_Differences_Mining_Contrast_Sets

Bay, S. D., Pazzani, M. J.(bez dat.). Detecting Change in Categorical Data: Mining Contrast Sets. Preuzeto 11.09.2020. s https://www.ics.uci.edu/~pazzani/Publications/stucco.pdf

Boettcher, M. (2011). Contrast and change mining. Preuzeto 12.09.2020. s https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.27 , str. 1.

Contrast set learning. (bez dat.). U Wikipedia. Preuzeto 12.09.2020. s https://en.wikipedia.org/wiki/Contrast_set_learning

Dong, G., i Bailey, J. (2012). Contrast Data Mining: Concepts, Algorithms, and Applications. CRC Press. Taylor & Francis Group, 2012.

Dujman, J. (2017). Upotreba neparametarske statistike u inženjerskim problemima (Završni rad, Sveučilište u Zagrebu, Fakultet strojarstva i brodogradnje, Zagreb). Preuzeto s https://zir.nsk.hr/islandora/object/fsb%3A3729

Kralj, P., Lavrač, N., Gamberger et al. (2007). Contrast set mining through subgroup discovery applied to brain Ischaemina dana. Preuzeto 17.09.2020. s https://sci2s.ugr.es/keel/pdf/specific/congreso/PAKDD-2007-KraljEtAl-CSMthroughSD.pdf

Nazeri, Z., Donohue, G, Sherry, L. (2008). Analyzing Relationships Between Aircraft Accidents and Incidents A data mining approach. Preuzeto 17.09.2020. s http://www.icrat.org/icrat/seminarContent/2008/Analyzing_Relationships_Between_Aircraft_Accidents.pdf

Oreški, D. (2014). Evaluacija tehnika otkrivanja kontrasta za potrebe selekcije atributa radi klasifikacije (Doktorski rad, Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin). Preuzeto s https://dr.nsk.hr/islandora/object/foi%3A737

SAS Institute (bez dat.) What is data mining | SAS ? Preuzeto 11.09.2020. s https://www.sas.com/en_us/insights/analytics/data-mining.html

Webb, G. I. (1995). *OPUS: An Efficient Admissible Algorithm for Unordered Search.* Preuzeto 15.09.2020. s https://arxiv.org/pdf/cs/9512101.pdf

# 8. Popis slika

# 9. Popis tablica

# 10. Prilozi

Excel tablica vezana uz podatke vlastitog primjera primjene STUCCO algoritma, poslana je u prilogu prilikom slanja verzije u sustav.

Prilog 1. Izgled Excel tablice korištene u izradi vlastitog primjera

| | BreedIDDesc | GenderIDDesc | vaccination_yrs | WhereBittenIDDesc |
|---|---|---|---|---|
| 2 | BOXER | UNKNOWN | -1 | BODY |
| 3 | BOXER | UNKNOWN | -1 | HEAD |
| 4 | BOXER | UNKNOWN | -1 | BODY |
| 5 | BOXER | UNKNOWN | -1 | BODY |
| 6 | BOXER | UNKNOWN | -1 | BODY |
| 7 | BOXER | UNKNOWN | -1 | BODY |
| 8 | BOXER | UNKNOWN | -1 | UNKNOWN |
| 9 | BOXER | UNKNOWN | -1 | BODY |
| 10 | BOXER | UNKNOWN | -1 | BODY |
| 11 | BOXER | UNKNOWN | -1 | HEAD |
| 12 | BOXER | UNKNOWN | -1 | UNKNOWN |
| 13 | BOXER | UNKNOWN | -1 | BODY |
| 14 | BOXER | UNKNOWN | -1 | BODY |
| 15 | BOXER | UNKNOWN | -1 | BODY |
| 16 | BOXER | UNKNOWN | -1 | BODY |
| 17 | BOXER | UNKNOWN | -1 | BODY |
| 18 | BOXER | UNKNOWN | -1 | BODY |
| 19 | BOXER | UNKNOWN | -1 | UNKNOWN |
| 20 | BOXER | UNKNOWN | -1 | BODY |
| 21 | BOXER | UNKNOWN | -1 | BODY |
| 22 | BOXER | UNKNOWN | -1 | BODY |
| 23 | BOXER | UNKNOWN | -1 | BODY |
| 24 | BOXER | UNKNOWN | -1 | UNKNOWN |
| 25 | BOXER | UNKNOWN | -1 | BODY |
| 26 | BOXER | UNKNOWN | 1 | BODY |
| 27 | BOXER | UNKNOWN | 1 | BODY |
| 28 | BOXER | UNKNOWN | 1 | UNKNOWN |
| 29 | BOXER | UNKNOWN | 1 | BODY |
| 30 | BOXER | UNKNOWN | 3 | BODY |
| 31 | BOXER | UNKNOWN | 3 | BODY |
| 32 | BOXER | MALE | -1 | BODY |

GRUPE: LR (labrador), GS (njemački ovčar), CH (čivava), BO (bokser)

c: Gender = MALE & vaccination_yrs = -1 (muški psi koji nisu cijepljeni)

| | | | | |
|---|---|---|---|---|
| sup (c\|BO) = | 38/181= | 0.209945 | 20.99% | |
| sup (c\|CH) = | 47/165= | 0.284848 | 28.48% | max |
| sup (c\|GS) = | 66/327= | 0.201835 | 20.18% | |
| sup (c\|LR) = | 50/253= | 0.197628 | 19.76% | min |

max -min = 28,48 - 19,76 = 8.72%
set je *velik*

značajnost:

| stvarne: | BO | CH | GS | LR |
|---|---|---|---|---|
| c | 38 | 47 | 66 | 50 |
| ne c | 143 | 118 | 261 | 203 |

N = 181+165+327+253 926

| očekivane: | BO | CH | GS | LR |
|---|---|---|---|---|
| c | 39.28834 | 35.81533 | 70.97948 | 54.91685 |
| ne c | 141.7117 | 129.1847 | 256.0205 | 198.0832 |

| razlika: | BO | CH | GS | LR |
|---|---|---|---|---|
| c | -1.28834 | 11.18467 | -4.97948 | -4.91685 |
| ne c | 1.288337 | -11.1847 | 4.979482 | 4.916847 |

| kvadrat razlike: | BO | CH | GS | LR |
|---|---|---|---|---|
| c | 1.659812 | 125.0967 | 24.79524 | 24.17538 |
| ne c | 1.659812 | 125.0967 | 24.79524 | 24.17538 |

| kv.r /očekivane: | BO | CH | GS | LR |
|---|---|---|---|---|
| c | 0.042247 | 3.492826 | 0.34933 | 0.440218 |
| ne c | 0.011713 | 0.968356 | 0.096849 | 0.122047 |

Prilog 2. Nastavak Excel tablice korištene u izradi vlastitog primjera

| | | | | |
|---|---|---|---|---|
| 33 | BOXER | MALE | -1 | BODY |
| 34 | BOXER | MALE | -1 | BODY |
| 35 | BOXER | MALE | -1 | BODY |
| 36 | BOXER | MALE | -1 | HEAD |
| 37 | BOXER | MALE | -1 | BODY |
| 38 | BOXER | MALE | -1 | HEAD |
| 39 | BOXER | MALE | -1 | BODY |
| 40 | BOXER | MALE | -1 | BODY |
| 41 | BOXER | MALE | -1 | BODY |
| 42 | BOXER | MALE | -1 | BODY |
| 43 | BOXER | MALE | -1 | BODY |
| 44 | BOXER | MALE | -1 | BODY |
| 45 | BOXER | MALE | -1 | HEAD |
| 46 | BOXER | MALE | -1 | UNKNOWN |
| 47 | BOXER | MALE | -1 | HEAD |
| 48 | BOXER | MALE | -1 | BODY |
| 49 | BOXER | MALE | -1 | BODY |
| 50 | BOXER | MALE | -1 | BODY |
| 51 | BOXER | MALE | -1 | HEAD |
| 52 | BOXER | MALE | -1 | HEAD |
| 53 | BOXER | MALE | -1 | BODY |
| 54 | BOXER | MALE | -1 | HEAD |
| 55 | BOXER | MALE | -1 | BODY |
| 56 | BOXER | MALE | -1 | BODY |
| 57 | BOXER | MALE | -1 | BODY |
| 58 | BOXER | MALE | -1 | BODY |
| 59 | BOXER | MALE | -1 | HEAD |
| 60 | BOXER | MALE | -1 | BODY |
| 61 | BOXER | MALE | -1 | BODY |
| 62 | BOXER | MALE | -1 | HEAD |
| 63 | BOXER | MALE | -1 | HEAD |
| 64 | BOXER | MALE | -1 | BODY |
| 65 | BOXER | MALE | -1 | BODY |
| 66 | BOXER | MALE | -1 | HEAD |
| 67 | BOXER | MALE | -1 | HEAD |
| 68 | BOXER | MALE | -1 | HEAD |
| 69 | BOXER | MALE | -1 | BODY |
| 70 | BOXER | MALE | 1 | HEAD |
| 71 | BOXER | MALE | 1 | BODY |
| 72 | BOXER | MALE | 1 | HEAD |
| 73 | BOXER | MALE | 1 | HEAD |
| 74 | BOXER | MALE | 1 | BODY |
| 75 | BOXER | MALE | 1 | BODY |

| | | | |
|---|---|---|---|
| hi kvadrat = | 5.52358 | > 5% | set *nije signifikantan* |

Prilog 3. Kraj tablice

| | | | | |
|---|---|---|---|---|
| 882 | LABRADOR RETRIV | FEMALE | -1 | BODY |
| 883 | LABRADOR RETRIV | FEMALE | -1 | BODY |
| 884 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 885 | LABRADOR RETRIV | FEMALE | 1 | HEAD |
| 886 | LABRADOR RETRIV | FEMALE | 1 | UNKNOWN |
| 887 | LABRADOR RETRIV | FEMALE | 1 | UNKNOWN |
| 888 | LABRADOR RETRIV | FEMALE | 1 | UNKNOWN |
| 889 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 890 | LABRADOR RETRIV | FEMALE | 1 | UNKNOWN |
| 891 | LABRADOR RETRIV | FEMALE | 1 | HEAD |
| 892 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 893 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 894 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 895 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 896 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 897 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 898 | LABRADOR RETRIV | FEMALE | 1 | UNKNOWN |
| 899 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 900 | LABRADOR RETRIV | FEMALE | 1 | UNKNOWN |
| 901 | LABRADOR RETRIV | FEMALE | 1 | HEAD |
| 902 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 903 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 904 | LABRADOR RETRIV | FEMALE | 1 | UNKNOWN |
| 905 | LABRADOR RETRIV | FEMALE | 1 | UNKNOWN |
| 906 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 907 | LABRADOR RETRIV | FEMALE | 1 | HEAD |
| 908 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 909 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 910 | LABRADOR RETRIV | FEMALE | 1 | HEAD |
| 911 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 912 | LABRADOR RETRIV | FEMALE | 1 | BODY |
| 913 | LABRADOR RETRIV | FEMALE | 3 | BODY |
| 914 | LABRADOR RETRIV | FEMALE | 3 | BODY |
| 915 | LABRADOR RETRIV | FEMALE | 3 | UNKNOWN |
| 916 | LABRADOR RETRIV | FEMALE | 3 | BODY |
| 917 | LABRADOR RETRIV | FEMALE | 3 | HEAD |
| 918 | LABRADOR RETRIV | FEMALE | 3 | UNKNOWN |
| 919 | LABRADOR RETRIV | FEMALE | 3 | UNKNOWN |
| 920 | LABRADOR RETRIV | FEMALE | 3 | UNKNOWN |
| 921 | LABRADOR RETRIV | FEMALE | 3 | BODY |
| 922 | LABRADOR RETRIV | FEMALE | 3 | UNKNOWN |
| 923 | LABRADOR RETRIV | FEMALE | 3 | HEAD |
| 924 | LABRADOR RETRIV | FEMALE | 3 | UNKNOWN |
| 925 | LABRADOR RETRIV | FEMALE | 3 | BODY |
| 926 | LABRADOR RETRIV | FEMALE | 3 | HEAD |
| 927 | LABRADOR RETRIV | FEMALE | 3 | BODY |