Analiza socio-demografskih čimbenika dijabetesa korištenjem otkrivanja znanja iz baza podataka

Burić, Marija

Sveučilište u Zagrebu

Ekonomski fakultet

Integrirani prijediplomski i diplomski sveučilišni studij

Poslovna ekonomija - smjer Menadžerska informatika

ANALIZA SOCIO-DEMOGRAFSKIH ČIMBENIKA DIJABETESA KORIŠTENJEM OTKRIVANJA ZNANJA IZ BAZA PODATAKA

Diplomski rad

Marija Burić

Zagreb, lipanj 2024.

Sveučilište u Zagrebu

Ekonomski fakultet

Integrirani prijediplomski i diplomski sveučilišni studij

Poslovna ekonomija - smjer Menadžerska informatika

ANALIZA SOCIO-DEMOGRAFSKIH ČIMBENIKA DIJABETESA KORIŠTENJEM OTKRIVANJA ZNANJA IZ BAZA PODATAKA

ANALYSIS OF SOCIO-DEMOGRAPHIC FACTORS OF DIABETES USING KNOWLEDGE DISCOVERY FROM DATABASES

Diplomski rad

Student: Marija Burić

JMBAG: 0067582434

Mentor: prof. dr. sc. Mirjana Pejić Bach

Zagreb, lipanj 2024.

IZJAVA O AKADEMSKOJ ČESTITOSTI

Izjavljujem i svojim potpisom potvrđujem da je diplomski rad / seminarski rad / prijava teme diplomskog rada isključivo rezultat mog vlastitog rada koji se temelji na mojim istraživanjima i oslanja se na objavljenu literaturu, a što pokazuju korištene bilješke i bibliografija.

Izjavljujem da nijedan dio rada / prijave teme nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog izvora te da nijedan dio rada / prijave teme ne krši bilo čija autorska prava.

Izjavljujem, također, da nijedan dio rada / prijave teme nije iskorišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

(vlastoručni potpis studenta)

(mjesto i datum)

STATEMENT ON THE ACADEMIC INTEGRITY

I hereby declare and confirm by my signature that the final thesis is the sole result of my own work based on my research and relies on the published literature, as shown in the listed notes and bibliography.

I declare that no part of the thesis has been written in an unauthorized manner, i.e., it is not transcribed from the non-cited work, and that no part of the thesis infringes any of the copyrights.

I also declare that no part of the thesis has been used for any other work in any other higher education, scientific or educational institution.

(personal signature of the student)

(place and date)

SAŽETAK

Dijabetes je kronična bolest koju karakteriziraju povišene razine glukoze u krvi. U današnje vrijeme, bolest postaje sve češća, a očekuje se kako će broj oboljelih i dalje značajno rasti. Rastuća rasprostranjenost bolesti uzrokovana je rastućim stopama pretilosti, sjedilačkim načinom života i lošim prehrambenim navikama što ga čini ozbiljnim javnozdravstvenim problemom u modernom svijetu. Otkrivanje znanja u bazama podataka predstavlja iznimno korisnu strategiju u borbi protiv dijabetesa. Analizom podataka o pacijentima, genetskih podataka i podataka o načinu života, istraživači mogu otkriti nove uvide u uzroke i progresiju bolesti. Otkriveno znanje pomaže u razvoju učinkovitih strategija prevencije, poboljšanju rane dijagnoze i prilagođavanju terapije pojedinačnim pacijentima, u konačnici poboljšavajući kvalitetu života oboljelih i smanjujući teret bolesti na zdravstvene sustave.

Ključne riječi: dijabetes, kvaliteta života, otkrivanje znanja iz baza podataka, klaster analiza

ABSTRACT

Diabetes is a chronic disease characterized by elevated blood glucose levels. Nowadays, the disease is becoming more and more common, and it is expected that the number of patients will continue to grow significantly. The growing prevalence of the disease is caused by rising rates of obesity, sedentary lifestyles and poor eating habits, making it a serious public health problem in the modern world. Knowledge discovery in databases is an extremely useful strategy in the fight against diabetes. By analysing patient data, genetic data, and lifestyle data, researches can uncover new insights into the causes and progression of disease. Discovered knowledge helps in developing effective prevention strategies, improving early diagnosis and tailoring therapy to individual patients, ultimately improving the quality of life od patients and reducing the burden of disease on healthcare systems.

Keywords: diabetes, life quality, knowledge discovery from databases, cluster analysis

SADRŽAJ

1. UVOD

1.1. Predmet i cilj rada

Predmet ovog rada je analiza socio-demografskih čimbenika dijabetesa putem otkrivanja znanja iz baze podataka. Otkrivanje znanja u bazama podataka je metodologija kojom se otkrivaju vrijedni podaci u bazama podataka, a glavni cilj je pronalazak korisnih informacija iz velikih količina podataka. Otkrivanje znanja iz baze podataka koja se sastoji od čimbenika povezanih s dijabetesom odnosi se na proces analize obilježja, veza i uzoraka unutar podataka kako bi se identificirali ključni čimbenici rizika, uzorci navika i ponašanja ili indikatori razvoja dijabetesa.

Cilj rada je pružiti sažeti pregled cijelog procesa i metoda za otkrivanje znanja iz baza podataka te primijeniti jednu od metoda otkrivanja znanja na socio-demografske čimbenike dijabetesa u svrhu identificiranja ključnih faktora rizika te njihov utjecaj na pojavu i razvoj bolesti.

Dodatno, cilj je dati analizu dobivenih rezultata i pružanje preporuka za razvoj učinkovitih strategija prevencije i liječenja bolesti koje se temelje na rezultatima provedene analize.

1.2. Izvor podataka i metode prikupljanja

Teorijski dio rada temelji se na sekundarnom istraživanju. Korišteni su različiti izvori podataka kao što su knjige, znanstveni članci vezani uz otkrivanje znanja iz baza podataka i dijabetes, web stranice i akademski članci pronađeni istraživanjem znanstvenih i stručnih časopisa.

Baza na temelju koje će se raditi obrada podataka u ovom radu je „Diabetes Health Indicators Dataset“[1], a pronađena je na web stranici „Kaggle“. Kaggle je platforma koja omogućuje istraživanje, analizu i otkrivanje znanja u raznovrsnim bazama podataka. Analiza će se provesti korištenjem softvera Weka. Weka je softver za analizu podataka i strojno učenje koji se koristi za istraživanje, razvoj i primjenu algoritama za obradu podataka. Metoda otkrivanja znanja u bazama podataka koja će se koristiti u ovom radu je klaster analiza.

[1] https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

1.3. Sadržaj i struktura rada

Rad se sastoji od pet glavnih dijelova, odnosno poglavlja. Osim uvoda i zaključka, rad obuhvaća i sljedeće dijelove: socio-demografske čimbenike dijabetesa, otkrivanje znanja u bazama podataka te analizu čimbenika dijabetesa korištenjem otkrivanja znanja iz baza podataka.

Uvod opisuje predmet i cilj rada, izvore podataka i metode prikupljanja. Prvi dio nakon uvoda pruža objašnjenje osnovnih pojmova o bolesti dijabetesa, njegovim različitim vrstama i uzrocima, daje pregled faktora rizika i uzroka s aspekta socio-demografije te opisuje moguće načine prevencije bolesti u kontekstu sociodemografskih okolnosti.

Treće poglavlje daje uvod u otkrivanje znanja iz baza podataka, opisuje cijeli proces i korake koje je potrebno provesti. Nadalje, objašnjava važnost primjene otkrivanja znanja iz baza podataka i njegovu ulogu u zdravstvenom sustavu te daje pregled najkorištenijih metoda za rudarenje podataka.

Glavni dio rada predstavlja analiza čimbenika dijabetesa korištenjem otkrivanja znanja iz baza podataka. Poglavlje obuhvaća metodologiju te rezultate istraživanja. Također, opisani su rezultati provedene analize te su na temelju provedene analize dane preporuke za razvoj strategija liječenja i prevencije bolesti.

Na kraju je dan zaključak u kojem je sumiran cijeli rad i naglašena važnost primjene otkrivanja znanja iz baza podataka u zdravstvu.

2. SOCIO-DEMOGRAFSKI ČIMBENICI DIJABETESA

2.1. Osnovni pojmovi dijabetesa

Dijabetes se definira kao skupina metaboličkih poremećaja te je jedan od najčešćih poremećaja koji pogađa više od pola bilijuna ljudi diljem svijeta. Procjenjuje se da će pojava dijabetesa dramatično porasti u nadolazećim godinama (Kavakiotis, 2017).

Dijabetes se može podijeliti na nekoliko različitih vrsta, a dva su glavna tipa, dijabetes tipa 1 i dijabetes tipa 2. Najčešći oblik je dijabetes tipa 2 koji ima 90% dijabetičara. U posljednjih nekoliko desetljeća rasprostranjenost dijabetesa tipa 2 značajno je porasla u zemljama svih razina dohotka (World Health Organization, n.d.). Glavni uzroci tipa 2 uključuju stil života, tjelesnu aktivnost, prehrambene navike i nasljeđe (Kavakiotis, 2017).

Dijabetes tipa 1 je kronično stanje, a za oboljele je pristup liječenju ključan za preživljavanje (World Health Organization, n.d.). Otprilike 5-10% ljudi koji imaju dijabetes imaju tip 1 koji se može dijagnosticirati u bilo kojoj dobi života, a simptomi se često razvijaju vrlo brzo (Centers for Disease Control and Prevention, 2023).

Globalno dogovoren cilj je zaustaviti porast dijabetesa i debljine do 2025. godine. Oko 422 milijuna ljudi diljem svijeta ima dijabetes, od kojih većina živi u zemljama s niskim i srednjim prihodima, a 1,5 milijuna smrti izravno se prepisuje dijabetesu svake godine. Broj slučajeva i rasprostranjenost dijabetesa u stalnom su porastu tijekom posljednjih nekoliko desetljeća (World Health Organization, n.d.).

Dijabetes tipa 2 predstavlja ozbiljan zdravstveni problem čija je pojava u porastu širom svijeta. Različiti faktori povezani sa stilom života i socio-demografskim karakteristikama doprinose visokom riziku od razvoja bolesti, uključujući dob, etnicitet, obiteljsku povijest, socioekonomski status, pretilost i nezdrave životne navike. U Europi, koja se suočava sa starenjem stanovništva, brojnim migrantima i etničkim manjinskim grupama te značajnom socioekonomskom raznolikošću unutar i među različitim zemljama, razumijevanje utjecaja tih čimbenika je od ključne važnosti za učinkovitu prevenciju i liječenje dijabetesa.

Neizmjenjivi čimbenici rizika kao što su dob, obiteljska povijest i etnicitet, zajedno sa socioekonomskim čimbenicima, ključni su za procjenu rizika i identifikaciju pojedinaca i skupina s visokim rizikom od dijabetesa. Ove skupine trebaju biti ciljane intenzivnim intervencijama i preventivnim programima. Primjerice, starija dob povećava rizik od razvoja dijabetesa tipa 2, a određene etničke skupine imaju genetsku predispoziciju za dijabetes što ih

čini ranjivijima na bolest. Socioekonomski status može utjecati na pristup zdravstvenoj skrbi i zdravoj prehrani čime se može dodatno povećati rizik od nastanka dijabetesa.

Razumijevanje socio-demografskih čimbenika nije ključno samo za identificiranje visokorizičnih pojedinaca, već i za oblikovanje ciljnih javnozdravstvenih intervencija koje uzimaju u obzir specifične potrebe različitih demografskih skupina. Ovakav pristup može poboljšati učinkovitost mjera prevencije i smanjiti teret dijabetesa na zdravstvene sustave na globalnoj razini.

S obzirom na to da je dijabetes povezan s čestom i preranom pojavom teških bolesti, od velike je važnosti što je ranije moguće uspostaviti dijagnozu bolesti i započeti s liječenjem. Ukoliko se bolest identificira u fazi predijabetesa, uspjesi terapije su puno veći (Kokić, 2009).

2.2. Faktori rizika i uzroci dijabetesa s aspekta socio-demografije

Činjenica da više od 80% osoba s dijabetesom tipa 2 ima prekomjernu tjelesnu masu ukazuje na čvrstu povezanost debljine i rizika obolijevanja od šećerne bolesti tipa 2. Na razvoj spomenutih stanja najviše utječe suvremeni način života karakteriziran prekomjernom konzumacijom industrijski prerađene hrane te sve niža razina tjelesne aktivnosti.

Rizik od nastanka dijabetesa je zavisan o kombinaciji genetskih faktora, životnih navika i drugih faktora. Faktori rizika kod dijabetesa tipa 1 nisu jasni kao kod predijabetesa i dijabetesa tipa 2. Poznati faktori rizika uključuju obiteljsku povijest i životnu dob. Rizik od nastanka je veći kod osoba koje imaju bliskog člana obitelji s dijabetesom tipa. Također, može se razviti u bilo kojoj životnoj dobi, ali najčešće se pojavljuje kod djece, tinejdžera i mladih osoba (Centers for Disease Control and Prevention, 2023).

Iako se rizični čimbenici dijabetesa tipa 1 još uvijek istražuju smatra se kako neki genetski i okolišni čimbenici, starija dob žene u vrijeme porođaja te izlaganje nekim bolestima povećavaju rizik od obolijevanja (Lakić, Džono-Boban, n.d). Vjeruje se da kombinacija genetske osjetljivosti i okolišnih čimbenika dovodi do dijabetesa tipa 1. Unatoč opsežnim istraživanjima mogućih uzroka, niti jedan još nije identificiran kao uzrok značajnog broja slučajeva izvan razumne sumnje (Roglic, 2016).

Faktori rizika od dijabetesa tipa 2 obuhvaćaju različite aspekte života koji povećavaju mogućnost nastanka ove metaboličke bolesti. Osobe s predijabetesom i povišenim raznima šećera u krvi posebno su podložne razvoju dijabetesa tipa 2. Nadalje, dob igra značajnu ulogu, rizik se povećava kod osoba starije dobi odnosno s navršenih 45 godina života, dok genetski

čimbenici odnosno obiteljska anamneza kod srodnika prvog stupnja, dodatno povećava vjerojatnost obolijevanja.

Nedostatak tjelesne aktivnosti također predstavlja značajan rizik jer fizička aktivnost pomaže u kontroli težine. Trudnice koje su razvile dijabetes za vrijeme trudnoće izlažu se povećanom riziku od dijabetesa kasnije tijekom života. Upravljanje navedenim faktorima rizika putem promicanja zdravih životnih navika, uključujući pravilnu prehranu, redovitu tjelesnu aktivnost te praćenje zdravstvenog stanja, ključni su za prevenciju ili odgodu razvoja dijabetesa tipa 2.

Niski socioekonomski status povezan je s većim rizikom za razvoj dijabetesa tipa 2. Osobe s nižim prihodima i obrazovanjem često imaju ograničen pristup kvalitetnoj zdravstvenoj skrbi, zdravoj prehrani i mogućnostima za fizičku aktivnost. Materijalna deprivacija, stres i nesigurni životni uvjeti dodatno doprinose povećanju rizika (Kyrou, I. i sur., 2020).

Starija dob se sve više prepoznaje kao važan čimbenik za razvoj dijabetesa tipa 2, što je posljedica produljenja životnog vijeka. Starenje povećava rizik od nastanka bolesti tako što narušava stvaranje inzulina i pojačava inzulinsku rezistenciju putem pretilosti i gubitka mišićne mase. Demografski trendovi starenja stanovništva, posebno u Europi, čine stariju dob ključnim faktorom za rastuću prevalenciju dijabetesa. Trenutno osobe u dobi od 65 godina i više imaju najvišu prevalenciju dijabetesa tipa 2 među svim dobnim skupinama, a očekuje se da će broj oboljelih u ovoj dobnoj skupini dodatno povećati. Podaci pokazuju kako srednje i visoko prihodovne zemlje imaju najvišu prevalenciju dijabetesa u dobnoj skupini od 60 do 74 godina i 75 do 79 godina, dok u nisko prihodovnim zemljama vrhunac prevalencije dijabetesa bilježi dobna skupina od 55 do 64 godina (Kyrou, I. i sur., 2020).

Etnicitet je nepromjenjivi čimbenik rizika za dijabetes tipa 2, s određenim etničkim skupinama koje imaju veći rizik bez obzira na prebivalište. Ove razlike proizlaze iz genetske predispozicije i veće osjetljivosti na kardiometaboličke komplikacije povezane s tjelesnom kompozicijom i pretilošću. Primjerice, osobe južnoazijskog, kineskog i japanskog podrijetla s prekomjernom težinom ili pretilošću imaju značajno veći rizik za dijabetes u usporedbi s Europljanima iste težine (Kyrou, I. i sur., 2020).

Nizak socioekonomski status je također jedan od čimbenika rizika povezan s nastankom dijabetesa tipa 2. Osobe s niskim socioekonomskim statusom imaju veću vjerojatnost razvoja dijabetesa tipa 2 zbog ograničenog pristupa kvalitetnoj zdravstvenoj skrbi, zdravoj prehrani i mogućnostima za tjelesnu aktivnost. Istraživanja pokazuju kako su u Europi ove nejednakosti posebno izražene među ženama (Kyrou, I. i sur., 2020).

Rizični faktori za nastanak dijabetesa u trudnoći predstavljaju različiti čimbenici uzrokovani životnim navikama i zdravstvenim stanjem. Što se tiče životne dobi, rizik od nastanka trudničkog dijabetesa se povećava već s navršenih 25 godina života, a genetski čimbenici i kod ovog tipa povećavaju vjerojatnost od nastanka. Ova vrsta dijabetesa obično nestaje nakon poroda, ali povećava rizik od dijabetesa tipa 2 (Centers for Disease Control and Prevention, 2022).

2.3. Prevencija bolesti u kontekstu sociodemografskih okolnosti

U današnje vrijeme broj oboljelih osoba raste alarmantnom brzinom što predstavlja ozbiljnu prijetnju zdravstvenom sustavu diljem svijeta. Uz porast stope pretilosti u dječjoj dobi, dijabetes postaje sve prisutniji među mladima, posebno među određenim etničkim skupinama (Harvard T.H. Chan, n.d.). Većina slučajeva predijabetesa i dijabetesa tipa 2 može se spriječiti promjenom načina života.

Prevencija dijabetesa tipa 2 može se postići kroz nekoliko jednostavnih koraka. Održavanje zdrave tjelesne težine je ključno jer prekomjerna težina predstavlja najvažniji čimbenik rizika za razvoj bolesti. Kod pretilih osoba, gubitak tjelesne težine može značajno smanjiti rizik od nastanka dijabetesa (Harvard T.H. Chan, n.d.).

Također, redovita tjelesna aktivnost važan je zaštitni čimbenik koji kod bolesti dijabetesa utječe na više načina. Program prevencije dijabetesa postavio je kao cilj umjerenu tjelesnu aktivnost. Navedeni cilj je odabran zbog izvedivosti i učinkovitosti u prevenciji dijabetesa, potkrijepljenog studijama koje pokazuju smanjen rizik od dijabetesa s povećanom tjelesnom aktivnošću (The Diabetes Prevention Program, 2002).

Promjene u načinu prehrane mogu imati veliki utjecaj na smanjenje rizika od obolijevanja. Važno je educirati o planiranju svakodnevnih obroka, a edukacija o zdravoj prehrani i izrada individualnih planova prehrane se preporuča za tip 1 i tip 2. Svim osobama s povećanim rizikom od šećerne bolesti treba omogućiti strukturirani edukacijski program (Vukić, Pravdić, 2020).

Istraživanje provedeno u Dijabetološkoj ambulanti imalo je za cilj saznati i prikazati važnost pravilne prehrane osoba oboljelih od dijabetesa. Istraživanje je pokazalo kako udruženo djelovanje nasljednih i čimbenika okoline (debljina, tjelesna neaktivnost, starija životna dob) imaju značajnu ulogu u nastanku dijabetesa tipa 2. Dokazano je kako kombinacija nepravilne prehrane, nedovoljne tjelesne aktivnosti i dugotrajan stres ubrzavaju vremenski tijek u pojavi bolesti. Dodatno, veliki broj studija ukazuje na vezu između konzumacije voća i povrća i

poboljšanja zdravlja, smanjenja rizika od određenih bolesti i usporavanja daljnjeg razvoja određenih bolesti, a među njima i dijabetesa (Vukić, Pravdić, 2020).

Dijabetes je jedan od brojnih zdravstvenih problema povezan s pušenjem. Pušači imaju 50% veću vjerojatnost razvijanja dijabetesa od nepušača. Pušenje utječe na nekoliko čimbenika koji mogu povećati otpornost na inzulin i ometaju djelovanje inzulina (Haire-Joshu i sur., 1999). Također, pušenje može povećati rizik od razvoja drugih zdravstvenih problema što može dodatno zakomplicirati kontrolu razine šećera u krvi. Prestanak pušenja je važan čimbenik u prevenciji dijabetesa i poboljšanju općeg stanja zdravlja.

Uključivanje socio-demografskih čimbenika koji utječu na nastanak bolesti u razvoj preventivnih mjera važno je za smanjenje prevalencije bolesti. Pri osmišljavanju preventivnih strategija u obzir treba uzeti socioekonomski status, dob, etnicitet, razinu obrazovanja i slično kako bi se razvile sveobuhvatne i učinkovite strategije. Stoga, preventivne mjere bi trebale uključivati politike koje olakšavaju pristup zdravoj prehrani i potiču tjelesnu aktivnost u zajednicama s nižim socioekonomskim statusom, primjerice putem subvencija za zdrave namirnice i izgradnje javnih parkova i fitness centara. Nadalje, potrebno je osigurati jezično i kulturno prilagođavanje zdravstvenih programa, suradnju s lokalnim zajednicama kako bi se promicale zdrave životne navike kod migranata i etničkih manjina. Također, škole i obrazovne institucije bi trebale imati važnu ulogu u promoviranju zdravog načina života od najranije dobi, čime se smanjuje rizik od razvoja dijabetesa kasnije u životu.

Iako su opsežna istraživanja dijabetesa pružila značajna saznanja tijekom proteklih desetljeća o genetskim ili okolišnim čimbenicima i staničnim mehanizmima, liječenju i upravljanju bolešću postoji još mnogo toga što treba otkriti i razjasniti. Kroz istraživanja dijagnoza, prognostička procjena odgovarajućeg liječenja i klinička administracija bi mogle steći značajnu prednost prema medicinskom rukovanju bolesti. U ovom nastojanju, oslanjanje na veliki i brzo rastući broj istraživanja i kliničkih podataka služi za uspostavljanje značajne osnove za sigurnu dijagnozu i daljnje liječenje. Stoga se rudarenje podataka i strojno učenje pojavljuju kao ključni procesi koji doprinose donošenju ispravnih odluka te se teži povezivanju podataka procjene do dijagnoze i donošenja odgovarajućih odluka (Kavakiotis, 2017).

3. OTKRIVANJE ZNANJA U BAZAMA PODATAKA

3.1. Uvod u otkrivanje znanja iz baza podataka

U današnje vrijeme obilje podataka predstavlja izazov i priliku te je stoga važno pronaći načine za automatsku analizu, klasifikaciju, sažimanje i otkrivanje važnih informacija u podacima. Brz rast podataka doveo je do potrebe za novim tehnikama i automatiziranim alatima koji mogu pomoći u pretvaranju ogromne količine podataka u korisne informacije i znanje. Navedeno je rezultiralo stvaranjem grane računalne znanosti koja se naziva rudarenjem podataka i njegove različite primjene. Rudarenje podataka ili otkrivanje znanja iz baza podataka automatizirano je izdvajanje uzoraka koji predstavljaju znanje pohranjeno ili pronađeno u velikim bazama podataka, skladištima podataka, webu, drugim masivnim spremištima informacija ili tokovima podataka (Han i sur., 2012).

Organizacije diljem svijeta svakodnevno generiraju ogromne količine podataka. Moćni i svestrani alati prijeko su potrebni za automatsko otkrivanje vrijednih informacija iz ogromnih količina podataka i za transformiranje takvih podataka u organizirano znanje. Ova potreba je dovela do rađanja rudarenja podataka, mladog, dinamičnog i perspektivnog područja.

Otkrivanje znanja u bazama podataka je proces pronalaženja vrijednih informacija u velikim bazama podataka. Cilj otkrivanja znanja iz baza podataka je pronalazak informacija kojima se može unaprijediti ili poboljšati poslovanje poduzeća (Pejić Bach, 2023). Proces otkrivanja znanja je iterativni niz koji se sastoji od čišćenja, integracije, selekcije, transformacije i rudarenja podataka te evaluacije uzoraka i predstavljanja znanja. Prvo se provodi čišćenje podataka kako bi se uklonili šumovi i nekonzistentni podaci, zatim se vrši integracija podataka ako se kombinira više izvora podataka. Nakon toga slijedi selekcija podataka relevantnih za analizu, a zatim transformacija podataka u prikladne oblike za rudarenje primjenom operacija sažimanja ili agregacije. Ključni korak je proces rudarenja podataka u kojem se primjenjuju metode za izvlačenje uzoraka. Zatim se provodi evaluacija uzoraka kako bi se identificirali uzorci koji predstavljaju znanje. Na kraju se znanje predstavlja korisnicima kroz tehnike vizualizacije i prezentacije znanja (Han, Kamber, Pei, 2012).

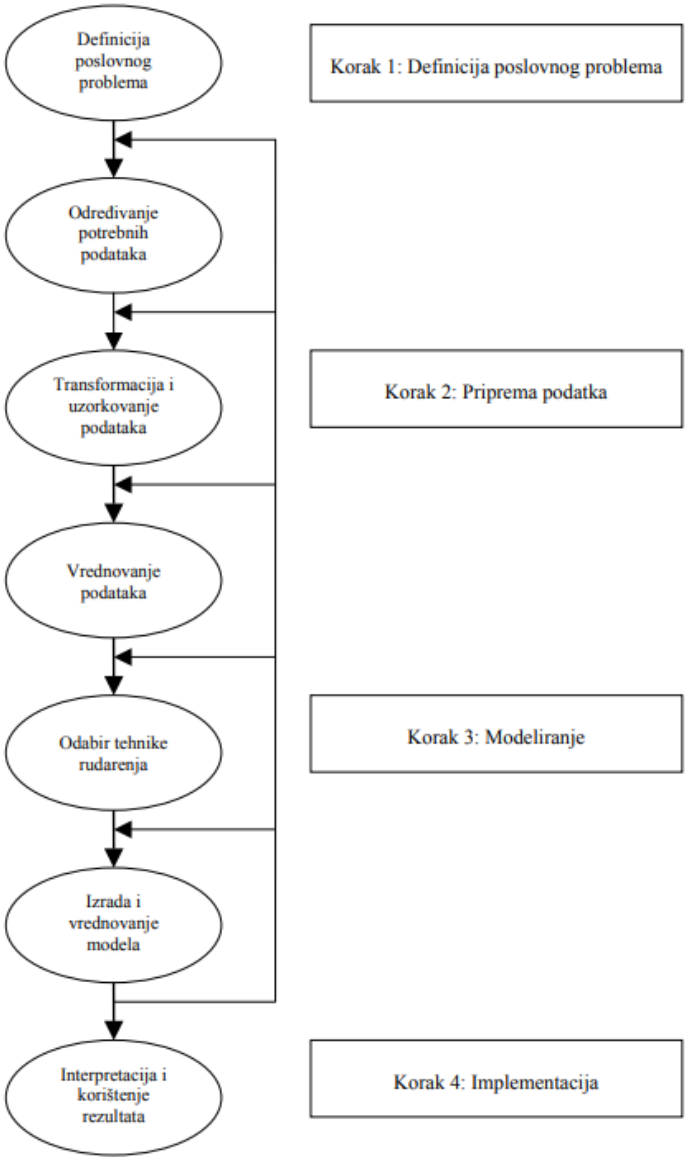
Izvori podataka mogu uključivati baze podataka, skladišta podataka, web, druga spremišta informacija ili podaci koji se dinamički prenose u sustav.

Rudarenje podataka predstavlja interdisciplinarno područje koje primjenjuje raznolike tehnike i alate radi automatskog otkrivanja složenih uzoraka, asocijacija, anomalija i struktura iz velikih količina podataka. Cilj rudarenja podataka je ekstrakcija vrijednih informacija iz ogromnih

skupova podataka pohranjenih u izvorima podataka. U suštini, rudarenje podataka ima za cilj istražiti i analizirati podatke u svrhu dubljeg razumijevanja sadržaja.

3.2. Proces otkrivanja znanja iz baza podataka

Proces otkrivanja znanja iz baza podataka se provodi kroz nekoliko različitih koraka: definicija poslovnog problema, priprema podataka, modeliranje i implementacija (Pejić Bach, 2007).



Slika 1 Prikaz procesa otkrivanja znanja

Izvor: Pejić Bach, M., & Kerep, I. (2011). Weka–alat za otkrivanje znanja iz baza podataka.

Prvi korak procesa otkrivanja znanja u bazama podataka odnosi se na definiciju poslovnog problema. Definiranje poslovnog problema ima za cilj utvrditi ciljeve analize podataka i razjasniti kako će provedena analiza pridonijeti rješavanju identificiranog problema.

U sklopu ovog koraka potrebno je pronaći kritično područje poslovanja, odrediti ciljeve projekta te odrediti članove tima. Potrebno je identificirati poslovni kontekst u kojem se primjenjuje otkrivanje znanja te specifičnosti područja poslovanja. Nakon toga je potrebno utvrditi ciljeve jer se jasnim poslovnim ciljevima osigurava da se analiza usredotoči na rješavanje specifičnih izazova ili prilika. Dodatno, određivanje ciljeva može pomoći u učinkovitom vođenju procesa analize podataka i maksimiziranju vrijednosti analize za organizaciju (Marbán i sur., 2009). Mogući ciljevi projekta su: analiza profila kupaca kojom se analiziraju zajedničke karakteristike ciljne populacije, segmentacija kojom se kupci nakon analize profila dijele u različite segmente, modeli odaziva koji predstavljaju procjenu vjerojatnosti da će kupac pozitivno odgovoriti na ponudu, procjena rizičnosti kupca, aktivacija odnosno procjena vjerojatnosti da će kupac početi koristiti proizvod, prodaja dodatnih proizvoda, odlazak kupca kod konkurencije, predviđanje profitabilnosti klijenta kroz određeni period (Pejić Bach, 2023).

Drugi korak se odnosi na pripremu podataka s ciljem osiguravanja prikladnosti podataka za postupke analize. Priprema podataka može oduzeti više vremena i predstavljati veći izazov od samog rudarenja jer podaci mogu biti nepotpuni, šumoviti i nedosljedni što može prekriti korisne obrasce. Dobra priprema podataka stvara set podataka koji je manji od originalnog što može značajno poboljšati učinkovitost rudarenja podataka. Dodatno, priprema podataka generira kvalitetne podatke što dovodi do opažanja kvalitetnih obrazaca. Iz navedenog se može zaključiti kako priprema podataka nije jednostavan i mali zadatak (Zhang i sur., 2003).

Priprema podataka se sastoji od različitih koraka. Prvo se odabire izvor podataka, a zatim se odabiru zavisne i nezavisne varijable. Zatim se provodi transformacija podataka koja obuhvaća pripremu podataka u tabličnom obliku, uključujući organizaciju podataka u redove (opažanja) i stupce (varijable) te primjenu operacija kao što su filtriranje, agregiranje i selekcija. Transformacija varijabli omogućuje izračunavanje novih varijabli kako bi se optimizirao proces analize podataka. Nadalje, uzorkovanje uključuje odabir odgovarajuće količine podataka za analizu, što može varirati ovisno o metodi analize i poslovnom problemu. Nakon toga se provodi vrednovanje podataka koje uključuje identifikaciju i tretiranje mogućih problema u podacima. Problemi mogu biti netipične vrijednosti (eng. *outliers*) i „prljave“ podatke kao što su nepostojeće vrijednosti, netočni podaci i nejasne definicije (Pejić Bach, 2023).

Treći korak predstavlja modeliranje koje uključuje primjenu različitih metoda za analizu podataka i izgradnju modela. Prvi dio se odnosi na izbor modela koji se odnosi na identifikaciju vrste problema koji se rješava te odabir odgovarajućeg modela ovisno o ciljevima. Nakon toga se podaci dijele te se izabiru značajke koje će se koristiti u modeliranju, a za odabrane stavke se provodi i optimizacija kako bi se što bolje prilagodilo podacima. Na kraju se izgrađuje model primjenom odabranog algoritma te evaluacija performansi korištenjem određenih mjera.

Metode modeliranja uključuju tehnike za otkrivanje grupa, predviđanje događaja i predviđanje vrijednosti. Modeliranje zahtijeva korištenje specijaliziranog softvera za analizu i obradu podataka. Metodama otkrivanja grupa se traže uzorci u podacima bez prethodnog znanja o njihovom obliku, a za to se koriste metode segmentiranja i asocijativnih pravila kojima se identificiraju grupe ili skupovi podataka koji dijele slične karakteristike ili ponašanja. Segmentacijske tehnike, kao što su algoritmi klasteriranja grupiraju podatke na temelju sličnosti, dok asocijativna pravila identificiraju veze između različitih varijabli ili atributa u podacima. Navedene metode omogućavaju dublje razumijevanje skrivenih uzoraka u podacima. Metode predviđanja događaja koriste se kako bi se procijenila vjerojatnost nekog događaja na temelju dostupnih podataka. Ove metode uključuju tehnike poput stabla odlučivanja, logističke regresije i neuronskih mreža. Metode za predviđanje vrijednosti koriste se za procjenu budućih vrijednosti na temelju podataka. Za ovakva predviđanja mogu se koristiti tehnike kao što su neuronske mreže, linearna regresija i metode vremenskih serija.

Vrednovanje rezultata uključuje procjenu uspješnosti modela kroz različite metrike, ovisno o vrsti analize koja se provodi (Pejić Bach, 2005).

Posljednji korak se odnosi na implementaciju rezultata, a postoje tri mogućnosti implementacije rezultata. Prva je implementacija gotovih indeksa koja generira model kako bi se informacije direktno primijenile u poslovnom okruženju. Druga mogućnost je izrada ad-hoc modela od strane internih stručnjaka ili vanjskih konzultanata što omogućava prilagođavanje modela specifičnim potrebama i uvjetima poslovanja. Treća mogućnost se odnosi na izradu alata koji podržavaju strateške, taktičke i operativne odluke. Ovo mogu biti softverski alati koji omogućuju praćenje performansi, predviđanje trendova ili automatizaciju poslovnih procesa na temelju analize podataka i rezultata modeliranja (Pejić Bach, 2023).

Proces modeliranja nikada nije u potpunosti završen, izrađeni modeli se kontinuirano nadograđuju i prilagođavaju novim podacima i promjenama kako bi se osigurala njihova relevantnost i korisnost.

3.3. Važnost primjene otkrivanja znanja iz baza podataka

Veliki napredak u biotehnologiji i zdravstvu doveo je do značajne proizvodnje podataka, kao što su genetski podaci i kliničke informacije, generirani iz elektroničkih zdravstvenih zapisa. Stoga je primjena metoda strojnog učenja i rudarenja podataka u bioznanostima neophodna kako bi se generirane informacije transformirale u korisno znanje. Opsežna istraživanja u svim aspektima dijabetesa dovela su do stvaranja velike količine podataka koje je potrebno iskoristiti za bolje razumijevanje bolesti u području istraživanja dijabetesa s obzirom na predviđanje i dijagnozu, komplikacije, genetsku pozadinu i zdravstvenu njegu (Kavakiotis i sur., 2017).

Kavakiotis i sur. (2017) smatraju kako je važno koristiti otkrivanje znanja iz baza podataka u istraživanju dijabetesa kako bi se omogućilo pronalaženje novih biomarkera koji bi mogli pomoći u predviđanju razvoja ili napredovanja bolesti. Za provođenje istraživanja je ključna dostupnost podataka, posebno onih koji se odnose na biološke karakteristike, kao što su genetske informacije ili podaci o biokemijskim procesima u tijelu. Putem integracije tehnika strojnog učenja i rudarenja podataka u skupove podataka koji sadrže kliničke i biološke informacije, istraživači mogu dobiti nove uvide u dijagnozu, simptome i liječenje dijabetesa (Kavakiotis i sur, 2017).

Al Yousef i sur. (2022) ističu ključnu ulogu rudarenja podataka u istraživanju dijabetesa, razjašnjavajući njegovu važnost u različitim domenama unutar područja. Tehnike rudarenja podataka i strojnog učenja korisne su u otkrivanju zamršenih odnosa među molekulama i stanjima i povezanosti lijekova i bolesti. Rudarenje podataka olakšava zadatke predviđanja iskorištavanjem biomarkera poput razine glukoze u krvi, omogućavajući rano otkrivanje i personalizirane strategije liječenja. Također, igra ključnu ulogu u istraživanju genetske pozadine i okolišnih čimbenika koji utječu na pojavu i razvoj dijabetesa. Integracija metoda rudarenja podataka i strojnog učenja obećava unaprjeđenje razumijevanja dijabetesa (Al Yousef i sur., 2022).

Rudarenje podataka može poboljšati različite segmente zdravstvenog sustava, nudeći prilike za predviđanje bolesti, unaprijeđeno liječenje i bolje upravljanje resursima. Rudarenje podataka omogućuje dijagnosticiranje i predviđanje bolesti analizom ogromnih količina podataka o pacijentima kako bi se identificirali obrasci i trendovi povezani s određenim bolestima. Ovakva mogućnost predviđanja iznimno je vrijedna tvrtkama za socijalno osiguranje radi mogućnosti predviđanje i učinkovite alokacije resursa za rješavanje budućih zdravstvenih potreba. Nadalje, rudarenje podataka olakšava klasifikaciju različitih bolnica na temelju njihove specijalizacije i

mogućnosti liječenja. Ovakva kategorizacija pomaže pacijentima odabir najprikladnije skrbi za njihovo zdravstveno stanje. Također, pomaže u procjeni učinkovitosti različitih načina liječenja analizom faktora kao što su uzroci, simptomi i troškovi liječenja. Ove informacije omogućuju pružateljima zdravstvenih usluga prilagodbu liječenja potrebama pacijenata. Dodatno, rudarenje podataka može pomoći u prevenciji infekcija u bolnicama identificiranjem potencijalnih izvora kontaminacije i omogućavanjem proaktivnih mjera za ublažavanje rizika. Naposljetku, rudarenje podataka olakšava identifikaciju visokorizičnih pacijenata, poput onih s kroničnim stanjima kao što je dijabetes (Rastogi, Bansal, 2023).

Analizom podataka o pacijentima, zdravstveni sustavi mogu implementirati preventivne mjere za učinkovitije upravljanje visokorizičnim pacijentima što u konačnici poboljšava ishode za pacijente i smanjuje ukupne troškove zdravstvene skrbi. Rudarenje podataka predstavlja vrijedan alat u sustavu zdravstvene skrbi, olakšava donošenje ispravnih odluka, optimizira raspodjelu resursa i pacijentima pruža kvalitetnu skrb.

3.4. Prikaz metoda za otkrivanje znanja iz baza podataka

Metode otkrivanja znanja iz baza podataka obuhvaćaju različite tehnike koje se koriste za identifikaciju uzoraka, informacija i znanja iz velikih baza podataka. Rudarenje podataka uključuje korištenje i obradu dostupnih podataka za donošenje prosudbi. U nastavku će se opisati neke od najkorištenijih metoda za rudarenje podataka.

Klasifikacija predstavlja metodu otkrivanja znanja iz baza podataka koja se koristi za kategoriziranje objekata baze podataka u unaprijed definirane kategorije ili klase na osnovu njihovih karakteristika. Navedena tehnika omogućava stvaranje modela koji „uči“ iz postojećih podataka kako bi mogao nove objekte klasificirati u odgovarajuće kategorije. Ovakav pristup se može koristiti kao korak pred obrade prije pohranjivanja podataka u model klasifikacije (Rastogi, Bansal, 2023). Cilj klasifikacije podataka je izraditi algoritam pomoću kojeg se s određenom vjerojatnošću može predvidjeti događanje jednog od mogućih ishoda (Pejić Bach, 2023).

Stablo odlučivanja je najkorišteniji klasifikacijski model, a služi kao statistički alat za obradu podataka putem grafičkog prikaza odluka i mogućih rezultata na osnovu ulaznih podataka (Rastogi, Bansal, 2023). Često je korištena metoda rudarenja podataka koja služi za uspostavljanje sustava klasifikacije na temelju višestrukih kovarijabli ili za razvoj algoritama previđanja za ciljnu varijablu. Metodologija stabla odlučivanja klasificira populaciju u segmente nalik granama koji grade obrnuto stablo s korijenskim čvorom, unutarnjim čvorovima

i listovima (Song, Lu, 2015). Ova metodologija je posebno korisna u rudarenju podataka i zadacima otkrivanja znanja jer nude jednostavan prikaz i logičnu strukturu što ih čini lakim za razumijevanje i interpretaciju. Zahtijevaju minimalnu pripremu podataka u usporedbi s drugim klasifikatorima te mogu rukovati različitim vrstama podataka i učinkovito upravljati nedostajućim vrijednostima. Također, imaju višestruke svrhe, uključujući zadatke klasifikacije i regresije što ih čini vrijednim alatom u različitim analitičkim kontekstima.

Metode segmentiranja koriste se radi rastavljanja skupa podataka u niz grupa odnosno klastera prema određenim karakteristikama uz uvjet da svaka grupa predstavlja homogen skup i uvjet razlikovanja svake grupe od ostalih grupa. Navedeno znači da su primjeri koji pripadaju nekoj grupi međusobno slični i da se primjeri koji pripadaju određenoj grupi značajno razlikuju od primjera drugih grupa. Glavni cilj segmentacije podataka je omogućiti fokusiraniju analizu i modeliranje za postizanje boljih rezultata. Različiti algoritmi, kao što su samo-organizirajuće neuralne mreže (Kohonen-ove mreže), probabilističke metode (AutoClass algoritam), algoritam k-srednjih vrijednosti (k-means) i joining tree clustering, mogu se koristiti ovisno o prirodi podatka i zadatku segmentacije. Na odabir algoritma utječu čimbenici poput računalne učinkovitosti, točnosti i specifičnih karakteristika podataka (Rastogi, Bansal, 2023).

Klasteriranje se koristi za identifikaciju skupova podataka koji dijele slične karakteristike, omogućujući njihovo grupiranje prema stupnju sličnosti što je važno za daljnje istraživanje i donošenje informiranih odluka (Rastogi, Bansal, 2023). Metode klasteriranja dijele primjere u klaster dok se istovremeno pridržavaju dva temeljna kriterija. Prvo, svaki klaster treba biti homogen, što znači kako atributi unutar istog klastera trebaju biti slični. Drugo, klasteri se trebaju međusobno razlikovati, osiguravajući da atributi u jednom klasteru nisi slični onima u drugim klasterima.

Algoritam K srednjih vrijednosti algoritam je nenadziranog strojnog učenja koji grupira neoznačeni skup podataka u različite klastere. Cilj grupiranja je podijeliti danu populaciju u K klastera odnosno više grupa tako da su podatkovne točke unutar svake grupe međusobno usporedive i drugačije od podatkovnih točaka unutar drugih grupa. Svaki klaster ima slične primjere, a atributi iz različitih klastera su međusobno različiti.

Asocijativna pravila su izjave koje pomažu u pokazivanju vjerojatnosti odnosa između stavki podataka unutar velikih skupova podataka u različitim bazama podataka. Rudarenje pravila povezivanja ima brojne primjene i široko je korištena metoda za pomoć u otkrivanju korelacija prodaje u transakcijskim podacima ili u skupovima medicinskih podataka. U znanosti o

podacima, asocijativna pravila koriste se za pronalaženje korisnih obrazaca i pravila u velikim skupovima podataka. Koriste se za objašnjenje uzoraka u podacima iz naizgled nezavisnih spremišta informacija, kao što su relacijske baze podataka i transakcijske baze podataka (Lutkevich, 2023).

Neuronske mreže ili umjetne neuronske mreže su napredni računalni modeli inspirirani biološkim sustavima, dizajnirani za otkrivanje obrazaca i stvaranje predviđanja (Jain i sur., 2013). Predstavljaju niz algoritama stvorenih kako bi prepoznali obrasce ili skrivene odnose podataka u skupu podataka. Sastoje se od ulaznog sloja, jednog ili više skrivenih slojeva i izlaznog sloja. Kada se u mreži nalazi jedan skriveni sloj radi se o plitkom učenju, a kod mreža sa više skrivenih slojeva radi se o dubokom učenju. Svi slojevi sastoje se od neurona, a svaki neuron prihvaća jedan ili više unosa te proizvodi određeni rezultat. Mogu se koristiti u različite svrhe, a neke od njih su računalni vid i operacije u stvarnom vremenu, prognoze, modeliranje i analiza, obrada prirodnog jezika, rudarenje podataka, rudarenje teksta i kupnja preko web-a (Pejić Bach, 2023). Posebno su učinkoviti u zadacima kao što su prepoznavanje uzoraka, donošenje odluka i predviđanje. Oponašaju mogućnosti obrade ljudskog mozga kroz međusobno povezane čvorove raspoređene u slojeve, gdje ulazne uzorke prima ulazni sloj, obrađuju kroz skrivene slojeve s ponderiranim vezama i na kraju proizvode rezultate putem izlaznog sloja (Rastogi, Bansal, 2023). Podaci koji ulaze u mrežu služe za treniranje mreže koja se pomoću utega, koji se nalaze između neurona svakog sloja, korigira za sve točnije rezultate.

4. ANALIZA ČIMBENIKA DIJABETESA KORIŠTENJEM OTKRIVANJA ZNANJA IZ BAZA PODATAKA

4.1. Metodologija istraživanja

Metodologija istraživanja temelji se na klaster analizi, a glavni cilj je grupiranje sudionika u slične skupine na temelju socio-demografskih čimbenika, navika i tjelesnog stanja kako bi se identificirali uzorci i povezanosti koji doprinose pojavi dijabetesa. Navedena analiza će omogućiti bolje razumijevanje čimbenika koji dovode do dijabetesa te razvijanje strategija usmjerenih na prevenciju i osvješćivanje o dijabetesu. Istraživanje će se provesti putem Weke, softvera za analizu podataka i strojno učenje koji se koristi za istraživanje, razvoj i primjenu algoritama za obradu podataka (Pejić Bach, Kerep, 2011).

Sustav nadzora faktora rizika u ponašanju (eng. *Behavioural Risk Factor Surveillance System* – BRFSS) je telefonska anketa o zdravlju koju godišnje prikuplja CDC. Anketa svake godine prikuplja odgovore više od 400 000 Amerikanaca o rizičnom ponašanju povezanom sa zdravljem, kroničnim zdravstvenim stanjima i korištenju preventivnih usluga. Korišteni skup podataka je očišćen i konsolidiran skup podataka, a izvorni skup podataka sadrži odgovore 441 455 osoba i ima 330 značajki. Značajke su pitanja koja se izravno postavljaju sudionicima ili izračunate varijable na temelju individualnih odgovora sudionika (Teboul, 2022).

Za izradu analize će se koristiti skup podataka „Diabetes Health Indicators Dataset" preuzet s Kaggle-a. U skupu podataka je sveukupno 253,680 primjera i 22 atributa pomoću kojih će se napraviti analiza. Baza podataka će se obraditi u Weka softveru. Weka je softverski alat otvorenog koda napisan u Java programskom jeziku, a razvijen je na Sveučilištu Waikato na Novom Zelandu. Weka korisnicima može pružiti podršku u cijelom procesu rudarenja podataka, uključujući pripremu podataka, statističku analizu shema učenja, vizualizaciju ulaznih podataka i rezultata učenja. Softver sadrži alate za klasifikaciju, regresiju, grupiranje, pripremu podataka i grafički prikaz asocijativnih pravila. Zahvaljujući Weki, korisnik primjenom tehnika strojnog učenja može iz baza podataka otkriti korisno znanje (Pejić Bach, Kerep, 2011).

Prvi korak u analizi se odnosi na pripremu podataka. Podaci su prikupljeni s Kaggle-a, a zatim su obrađeni u Notepadu i Excelu kako bi se podaci mogli transformirati u csv format koji će se moći analizirati u Weki.

U tablici 1 su prikazani atributi koji čine skup podataka, njihovi opisi, format atributa, modaliteti i najmanja i najveća vrijednost numeričkih atributa. Baza podataka sadrži 22 atributa, od kojih je 19 nominalnih, a 3 numerička. Svaki atribut ima svoj naziv, opis, format. Osim toga,

nominalni atributi imaju modalitete atributa, a numerički imaju prikazanu najmanju i najveću vrijednost numeričkih atributa.

Tablica 1 Popis atributa

Naziv atributa	**Opis atributa**	**Format atributa (numerički, binomni, nominalni)**	**Modaliteti atributa (nominalnih)**	**Najmanja i najveća vrijednost numeričkih atributa**
Diabetes_012	Dijabetes	Nominalni	No_diabetes, prediabetes, diabetes	
HighBP	Visoki krvni tlak	Nominalni	No_high BP, high_BP	
HighChol	Visoki kolesterol	Nominalni	No_high_cholesterol, high_cholesterol	
CholCheck	Provjera kolesterola u zadnjih 5 godina	Nominalni	no cholesterol check in 5 years, yes cholesterol check in 5 years	
BMI	Indeks tjelesne mase	Numerički	Body Mass Index	Min. 12 Max. 98
Smoker	Jeste li u životu popušili barem 100 cigareta?	Nominalni	No, yes	
Stroke	Jeste li imali moždani udar?	Nominalni	No, yes	
HeartDiseaseorAttack	Koronarna bolest ili infarkt miokarda	Nominalni	No, yes	

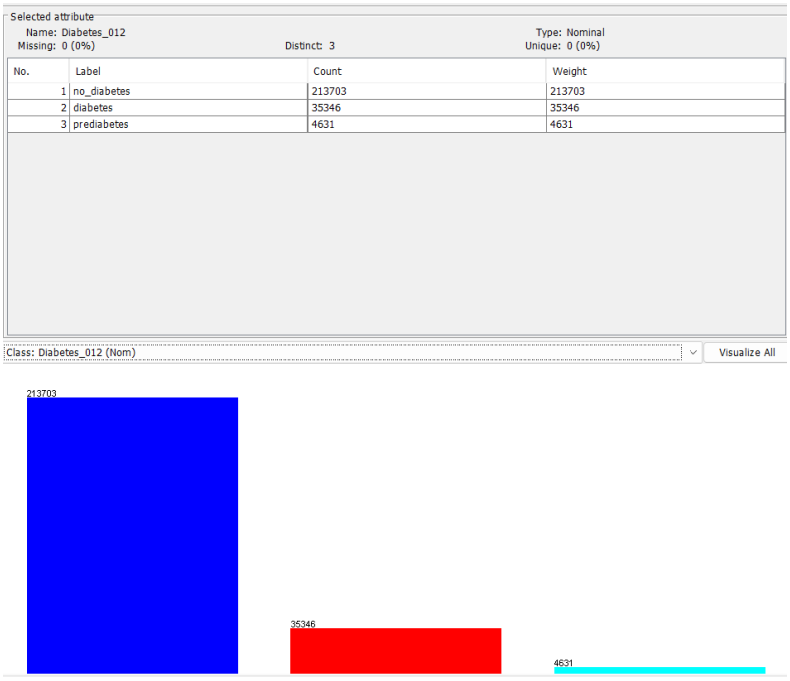
PhysActivity	Tjelesna aktivnost u zadnjih 30 dana	Nominalni	No, yes	
Fruits	Konzumiranje voća	Nominalni	No, yes	
Veggies	Konzumiranje povrća	Nominalni	No, yes	
HvyAlcoholConsump	Teška konzumacija alkohola (Odrasli muškarac >=14 pića tjedno/odrasla žena >=7 pića tjedno)	Nominalni	No, yes	
AnyHealthcare	Imate li bilo koji oblik zdravstvene zaštite?	Nominalni	No, yes	
NoDocbcCost	Jeste li u proteklih 12 mjeseci trebali posjetiti liječnika, ali niste mogli zbog troškova?	Nominalni	No, yes	
GenHlth	Općenito stanje zdravlja	Nominalni	Excellent, very good, good, fair, poor	

MentHlth	Koliko dana tijekom proteklih 30 dana mentalno zdravlje, uključujući stres, depresiju i probleme s emocijama, nije bilo dobro?	Numerički		Min. 1, Max. 30
PhysHlth	Koliko dana tijekom proteklih 30 dana fizičko zdravlje, uključujući fizičku bolest i ozljedu, nije bilo dobro?	Numerički		Min. 1, Max. 30
DiffWalk	Poteškoće s hodanjem ili penjanjem stepenicama	Nominalni	No, yes	
Sex	Spol	Nominalni	Female, male	
Age	Godine	Nominalni	18 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49, 50 to 54, 55 to 59, 60 to 64, 65 to 69, 70 to 74, 75 to 79, 80 or older	

Education	Razina obrazovanja	Nominalni	Never_attended_ school_or_only _kindergarten, Elementary, Some_ high_school, High_ school_graduate, Some_college_or_technical_ school, College_graduate	
Income	Visina prihoda	Nominalni	Less_than_$10000, Less_than_$15000, Less_than_$20000, Less_than_$25000, Less_than_$35000, Less_than_$50000, Less_ than_$75000, $75000_or_more	

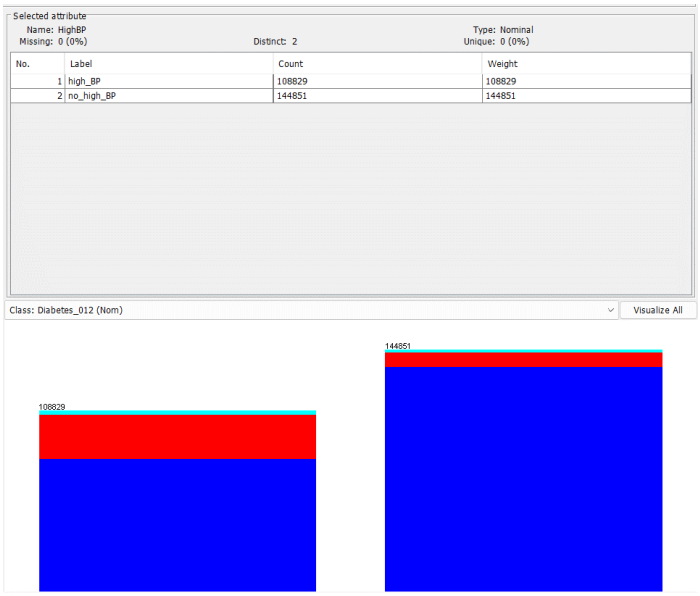
Izvor: Izrada autora

Na slici 2 prikazan je klasni atribut „Diabetes_012“ koji označava kategoriju dijabetesa te može poprimiti tri modaliteta – no_diabetes, prediabetes i diabetes. Modalitet no_diabetes vrijedi za 213703 (označeno plavom bojom), diabetes za 35346 (označeno crvenom bojom), a prediabetes za 4631 ispitanika (označeno svijetlo plavom bojom).



Slika 2 Prikaz klasnog atributa "Diabetes_012"

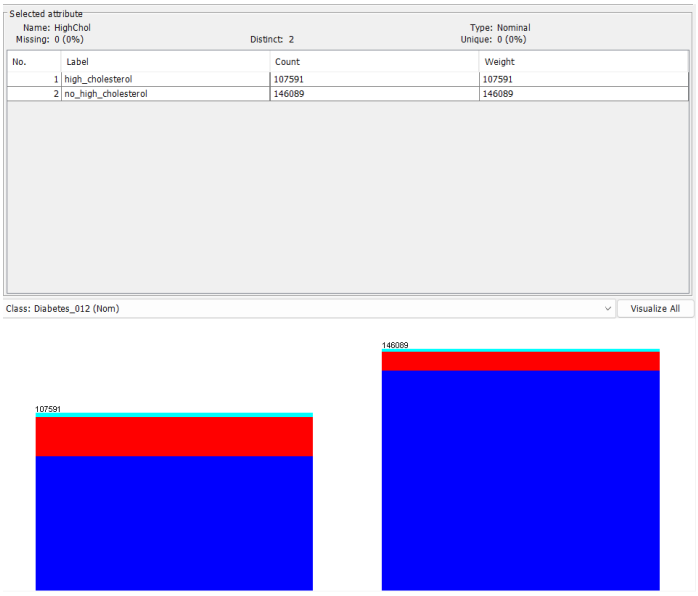
Atribut „HighBP“ pokazuje ima li ispitanik visoki krvni tlak. Atribut može poprimiti dva modaliteta – no_high_BP koji vrijedi za 144851 ispitanika i high_BP koji vrijedi za 108829 ispitanika. Pošto crvena boja označava ispitanike s dijabetesom može se zaključiti kako veći broj ispitanika s visokim krvnim tlakom ima dijabetes u odnosu na one bez visokog krvnog tlaka.



Slika 3 Atribut HighBP

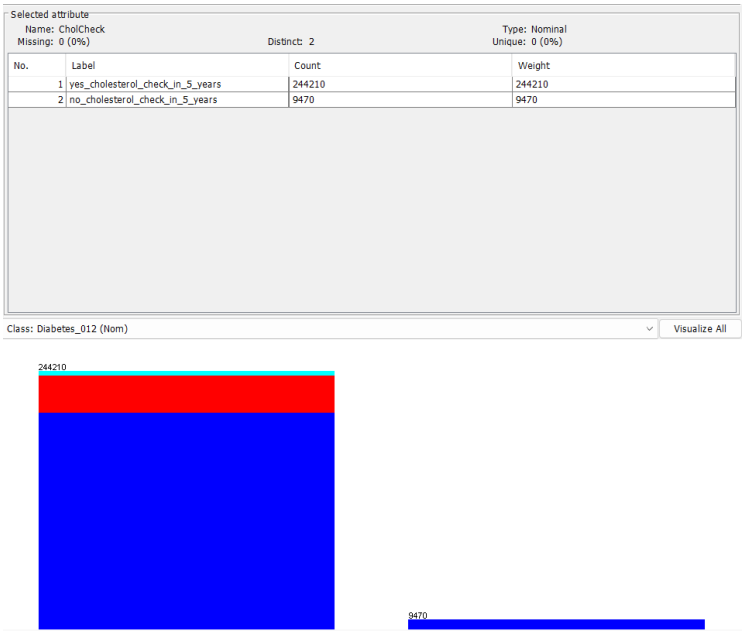
Atribut „HighChol“ govori ima li ispitanik visoki kolesterol. Modalitet high_cholesterol vrijedi za 107591 ispitanika, a modalitet no_high_cholesterol vrijedi za 146089 ispitanika. Vidljivo je

kako je kod osoba s visokim kolesterolom veći broj onih koji imaju dijabetes u odnosu na one bez visokog kolesterola.



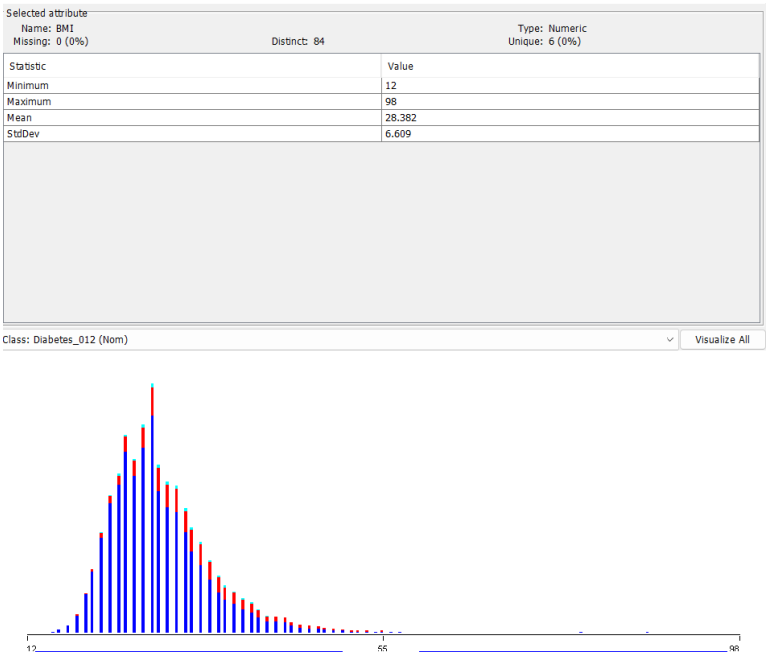
Slika 4 Atribut HighChol

Atribut „CholCheck“ pokazuje je li ispitanik u zadnjih 5 godina napravio provjeru kolesterola. 244210 ispitanika je obavilo provjeru, a 9470 ispitanika nije obavilo provjeru kolesterola.



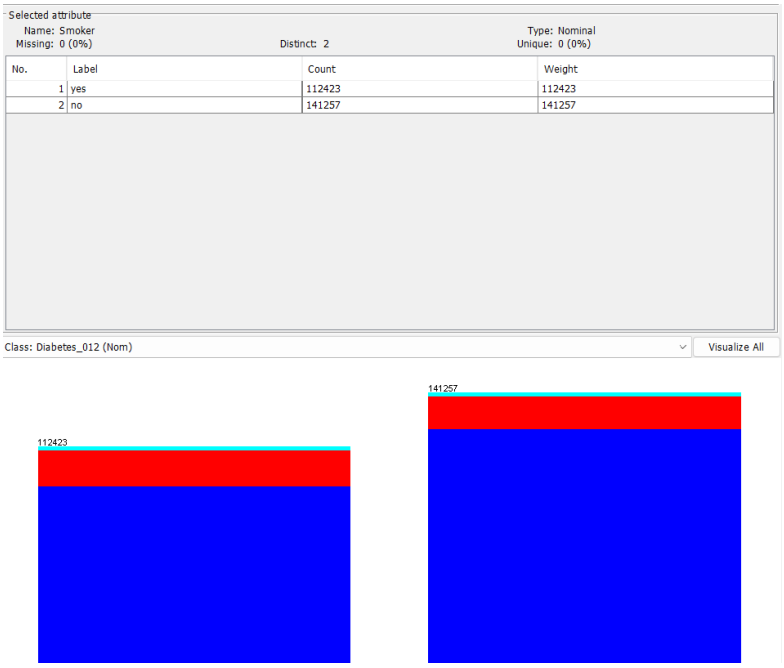
Slika 5 Atribut "CholCheck"

Atribut „BMI“, koji se odnosi na indeks tjelesne mase, kreće se u rasponu od 12 do 98. Indeks tjelesne mase je najuže povezan s količinom prekomjernog masnog tkiva u ljudskom tijelu. Iz prikazanog grafa se vidi kako je kod većeg BMI veći broj osoba s dijabetesom.



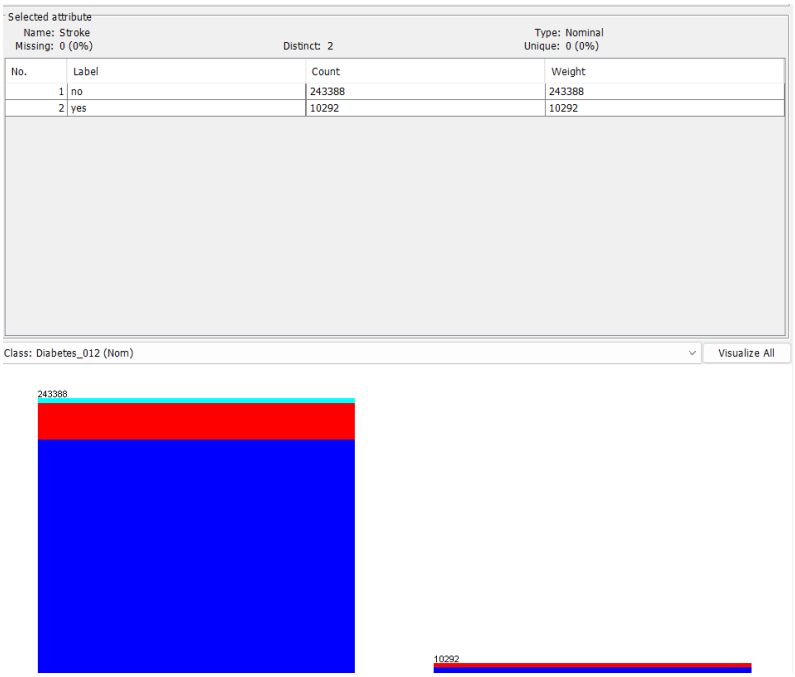
Slika 6 Atribut "BMI"

Atribut „Smoker“ odgovora na pitanje je li u životu konzumirano barem 100 cigareta. Modalitet yes se odnosi na 112423 ispitanika, a modalitet no na 141257 ispitanika.



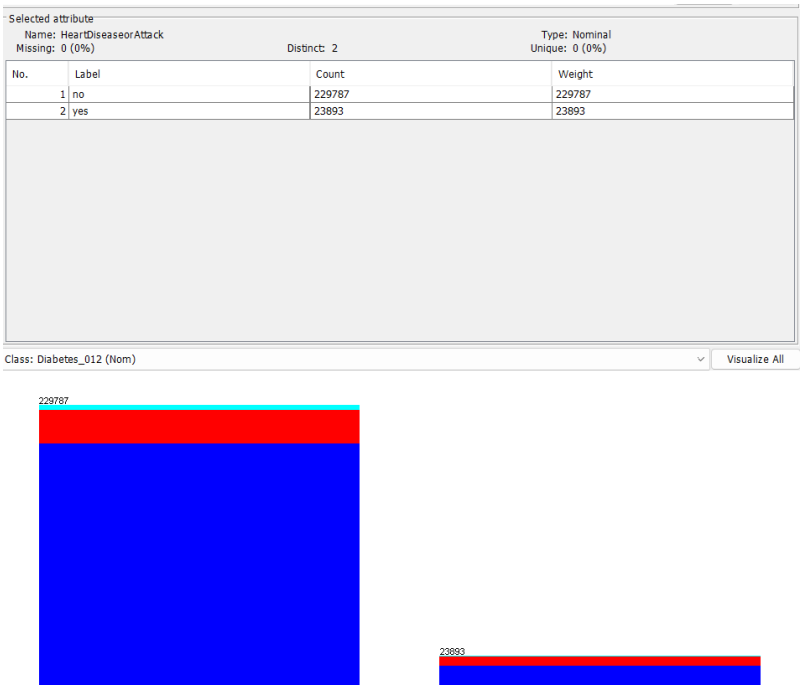
Slika 7 Atribut "Smoker"

Atribut „Stroke“ odgovora na pitanje je li ispitanik imao moždani udar. 243388 ispitanika je odgovorilo sa ne, a 10292 sa da.



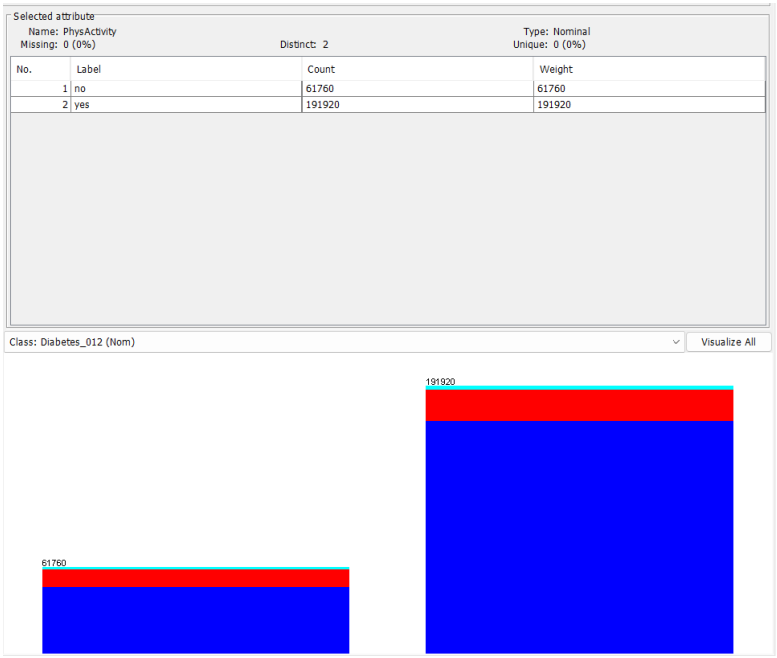
Slika 8 Atribut "Stroke"

Atribut „HeartDiseaseorAttack“ govori je li osoba oboljela od koronarne bolesti ili infarkta miokarda. 229787 ispitanika je odgovorilo sa ne dok je 23893 ispitanika odgovorilo sa yes.



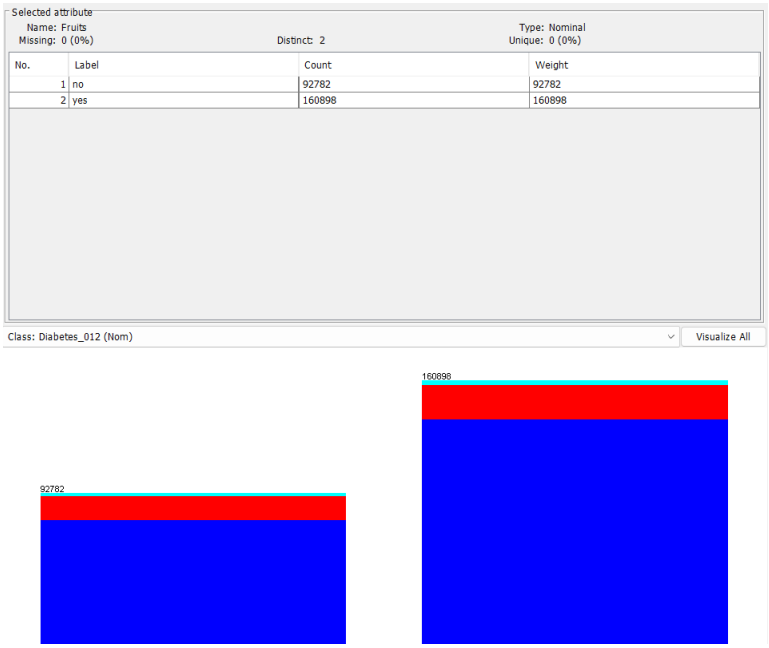
Slika 9 Atribut „HeartDiseaseorAttack“

Atribut „PhysActivity“ pokazuje je li ispitanik radio kakvu tjelesnu aktivnost u posljednjih 30 dana. 61760 ispitanika je odgovorilo s ne, a 19120 ispitanika je odgovorilo s da.



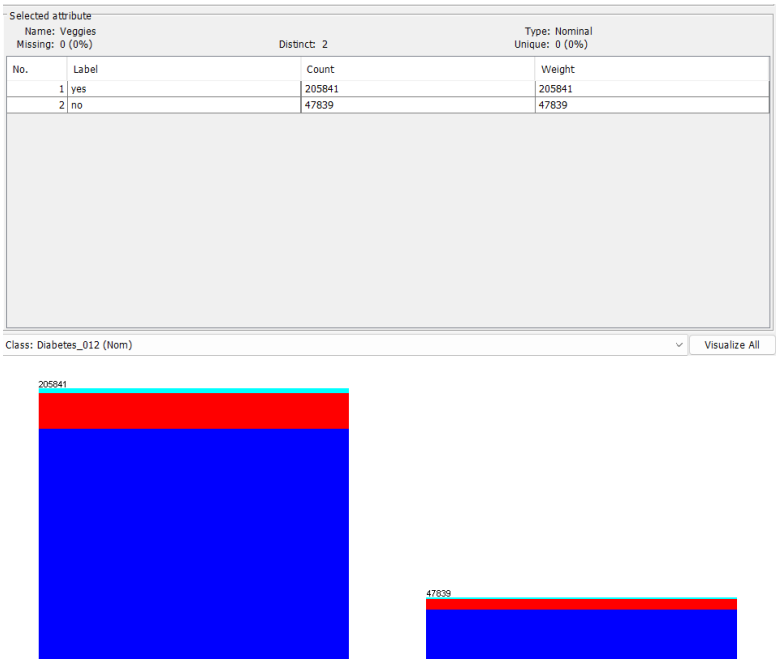
Slika 10 Atribut „PhysActivity“

Atribut „Fruits“ govori konzumira li osoba voće. 92782 ispitanika je odgovorilo s ne, a 160898 ispitanika je odgovorilo s da.



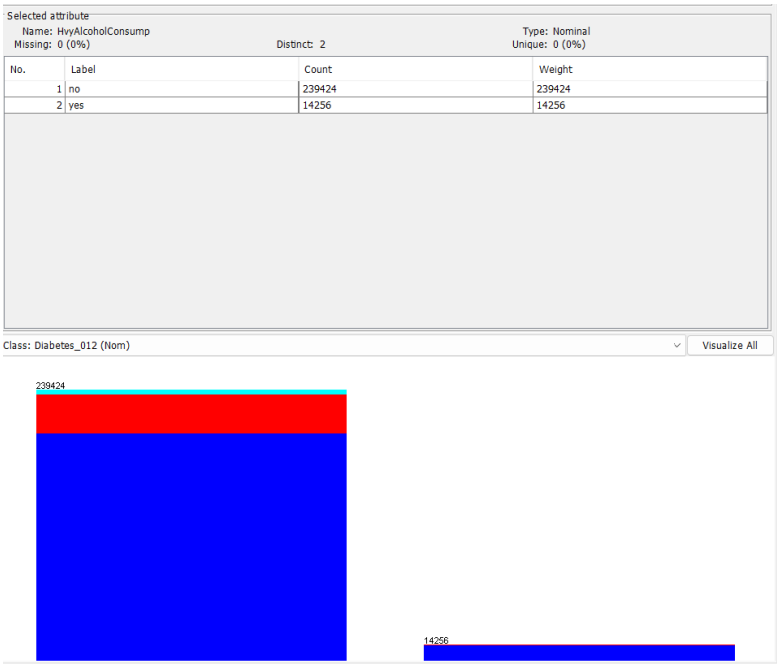
Slika 11 Atribut „Fruits“

Atribut „Veggies“ pokazuje konzumira li osoba povrće. 205841 ispitanika je odgovorilo s da, a 47839 ispitanika je odgovorilo ne.



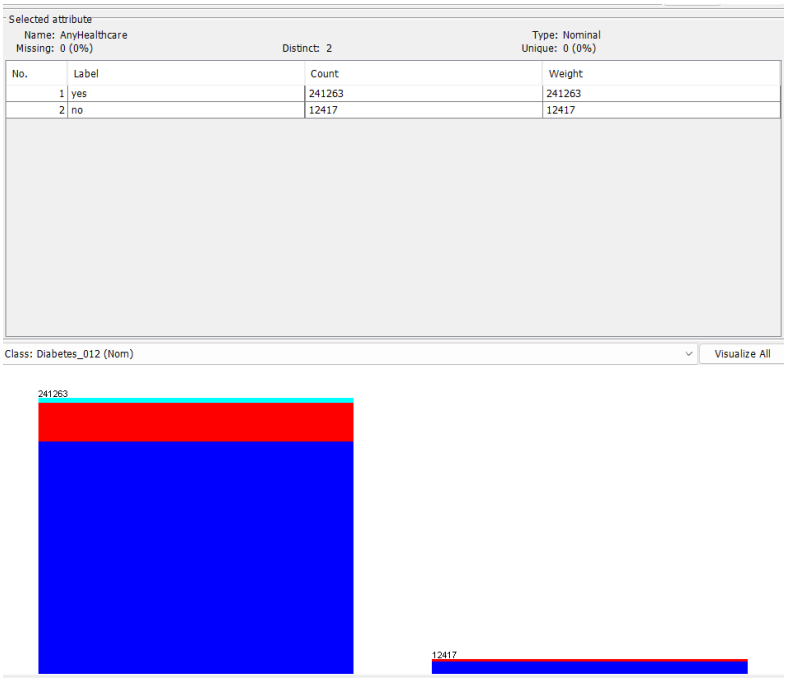
Slika 12 Atribut „Veggies“

Atribut „HvyAlcholConsump“ govori radi li se o teškoj konzumaciji alkohola. Teškom konzumacijom alkohola se smatra kada odrasli muškarac tjedno popije 14 ili više pića te kada odrasla žena popije 7 ili više pića tjedno. 239424 ispitanika je odgovorilo s ne, a 14256 ispitanika je odgovorilo s da.



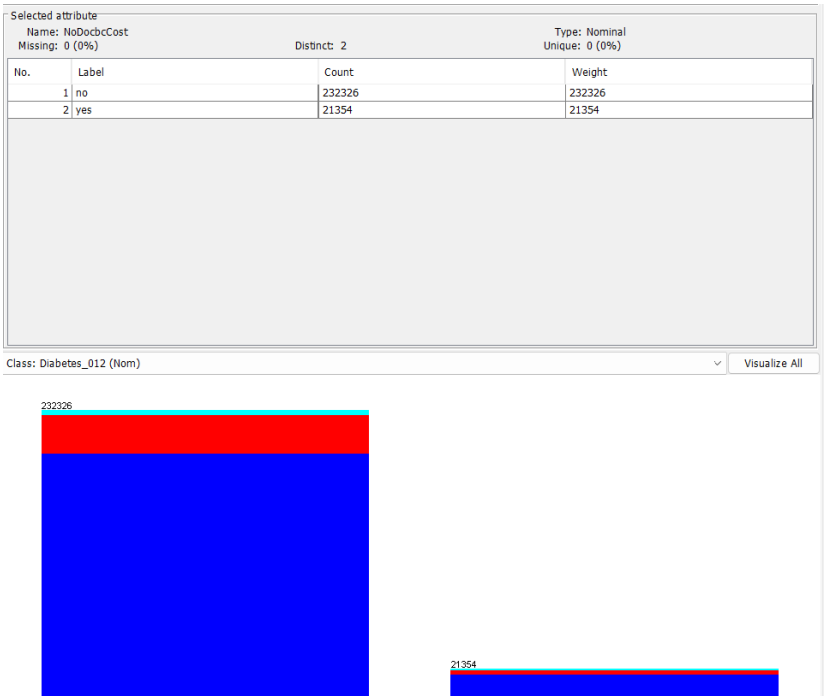
Slika 13 Atribut „HvyAlcholConsump“

Atribut „AnyHealthcare“ pokazuje ima li osoba bilo koji oblik zdravstvene zaštite. 241263 je odgovorilo s da, a 12417 ispitanika je odgovorilo s ne.



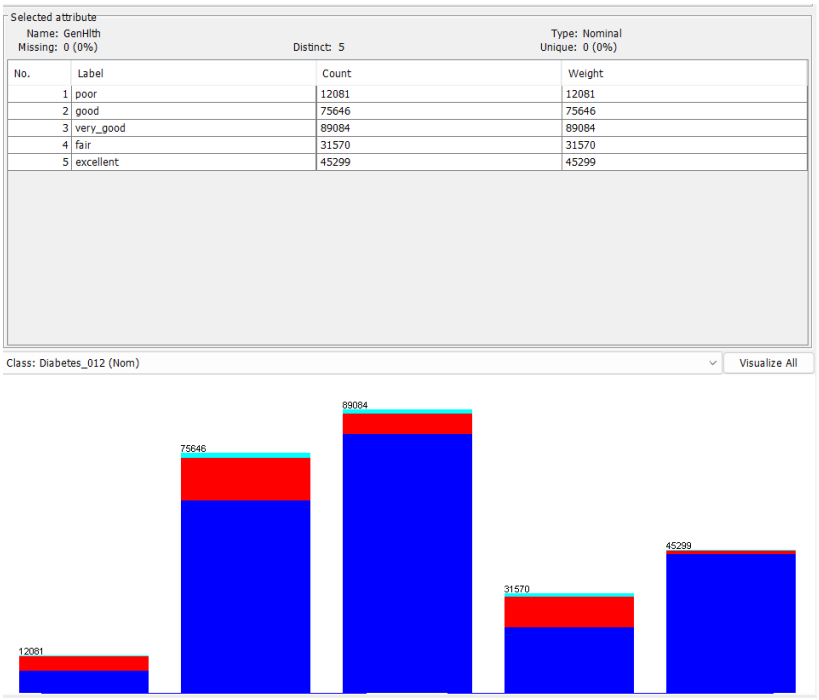
Slika 14 Atribut „AnyHealthcare“

Atribut „NoDocbcCost“ govori je li osoba u proteklih 12 mjeseci trebala posjetiti liječnika, ali nije mogla zbog troškova. 232326 ispitanika je odgovorilo s ne, a 21354 s da.



Slika 15 Atribut „NoDocbcCost“

Atribut „GenHlth“ govori kakvo je općenito stanje zdravlja. 12081 je odgovorilo „poor“, 75646 ispitanika je odgovorilo „good“, 89084 „very_good“, 31570 „fair“, a 45299 s „excellent“.

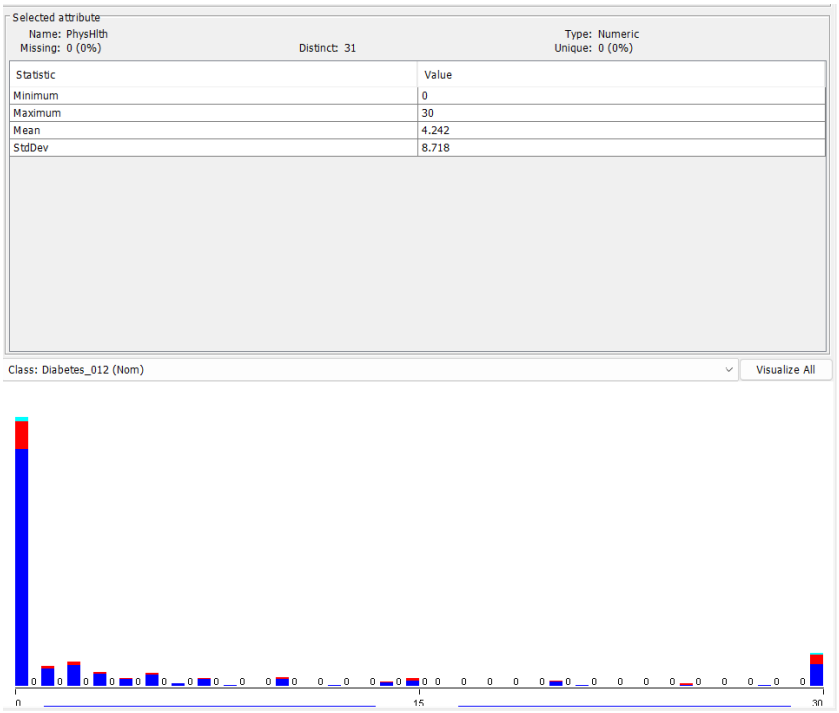


Slika 16 Atribut „GenHlth“

Atribut „MentHlth“ govori koliko dana tijekom proteklih 30 dana mentalno zdravlje, uključujući fizičku bolest i ozljedu, nije bilo dobro.

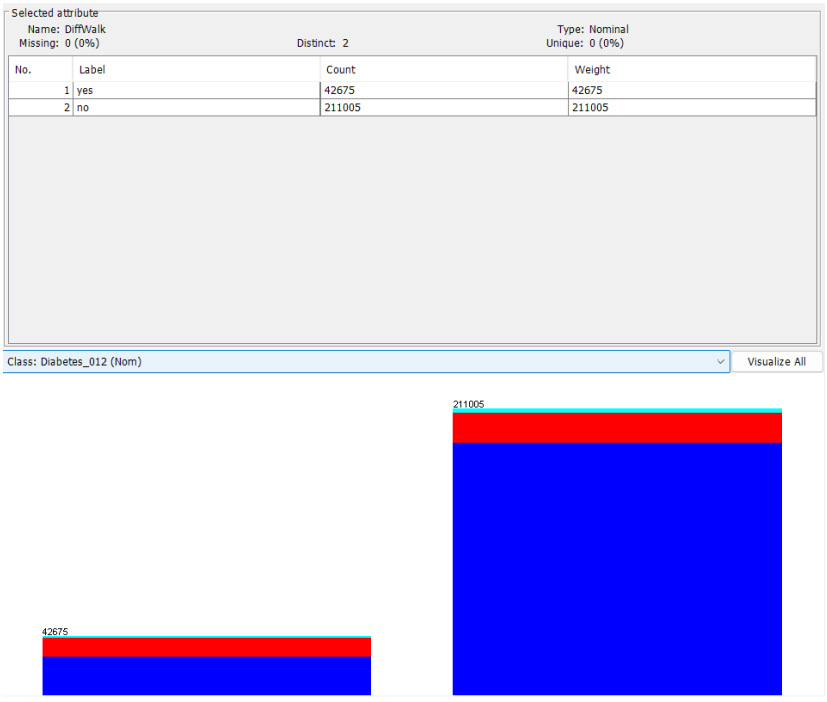
Slika 17 Atribut „MentHlth“

Atribut „PhysHlth“ govori koliko dana tijekom proteklih 30 dana fizičko zdravlje, uključujući fizičku bolest i ozljedu, nije bilo dobro.



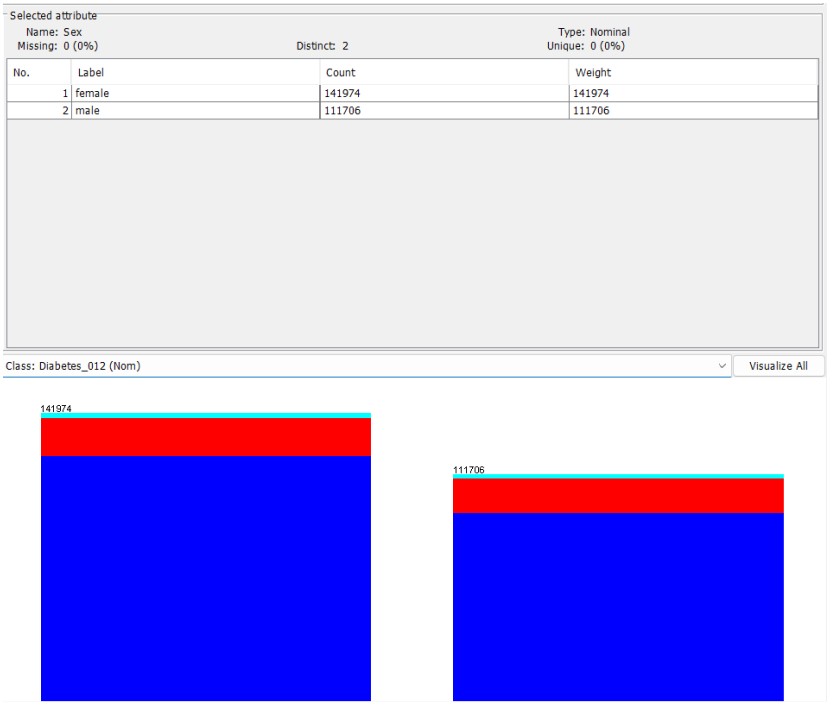
Slika 18 Atribut „PhysHlth“

Atribut „DiffWalk“ pokazuje ima li osoba poteškoće s hodanjem ili penjanjem stepenicama. 42675 ispitanika je odgovorilo s da, a 211005 je odgovorilo ne.



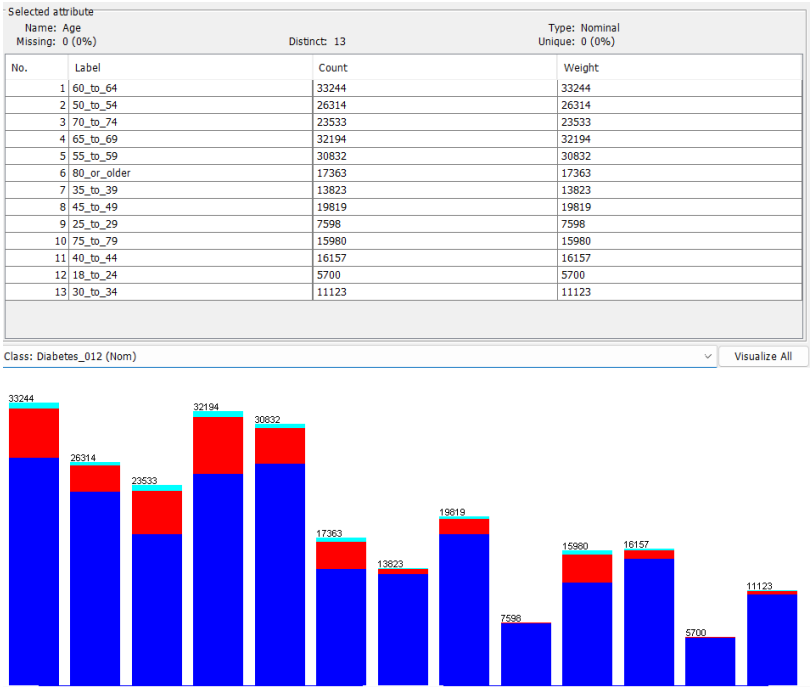
Slika 19 Atribut „DiffWalk“

Atribut „Sex“ pokazuje kojeg je spola osoba. 141974 ispitanika je ženskog, a 111706 ispitanika muškog spola.



Slika 20 Atribut „Sex“

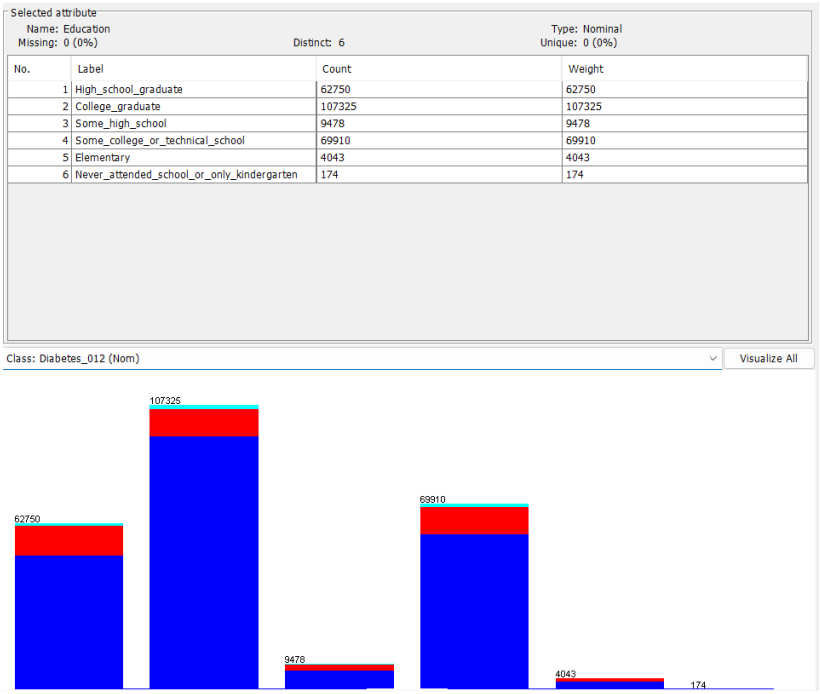
Atribut „Age“ pokazuje koliko godina ispitanici imaju te može poprimiti 13 modaliteta. Iz prikazanog grafa je vidljivo kako je dijabetes češći u starijoj dobi odnosno od 50 godine života.



Slika 21 Atribut „Age“

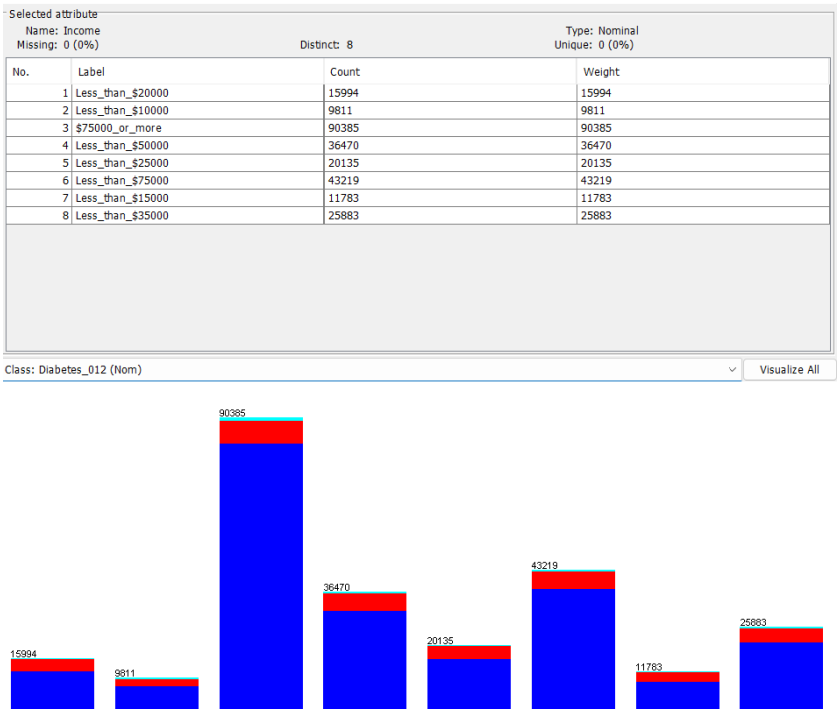
Atribut „Education“ označava razinu obrazovanja i može poprimiti 6 modaliteta. 62750 ispitanika ima „High_school_graduate“, 107325 je „College_graduate“, 9478 ima razinu

„Some_high_school“, 69910 ima „Some_college_or_technical_school“ razinu, 4043 ima „Elementary“ te 174 „Never_attended_school_or_only_kindergarten“ razinu.



Slika 22 Atribut „Education“

Atribut „Income“ označava visinu prihoda i ima 8 modaliteta. Less_than_$20000 vrijedi za 15994, less_than_$10000 za 9811, $75000_or_more za 90385, Less_than_$50000 za 36470, Less_than_$25000 za 20135, Less_than_$75000 za 43219, Less_than_$15000 za 11783 i Less_than_$35000 za 25883 ispitanika.



Slika 23 Atribut „Income“

4.2. Rezultati istraživanja

U nastavku će se provesti klaster analiza s ciljem istraživanja socio-demografskih čimbenika dijabetesa. Analiza će se fokusirati na 22 ključna čimbenika koji utječu na bolest dijabetesa. Glavni cilj ove analize je identificirati sličnosti među socio-demografskim čimbenicima, fizičkom stanju i životnim navikama ispitanika.

Za analizu provedenu u radu koristit će se korisničko sučelje Explorer. Sučelje Explorer pruža pristup svim objektima koristeći odabir menija i ispunjavanje formi odnosno obrazaca. Sastoji se od šest panela od kojih svaki služi za različite zadatke otkrivanja znanja iz baza podataka, a to su Preprocess, Classify, Cluster, Associate, Select attributes i Visualize panel (Pejić Bach, Kerep, 2011). Nakon odabira sučelja Explorer, otvara se panel Preprocess u koji je potrebno učitati uređeni set podataka u .csv formatu. U Preprocess panelu u Weki je vidljivo kako set podataka sadrži 22 atributa i 253680 instanci (primjera).

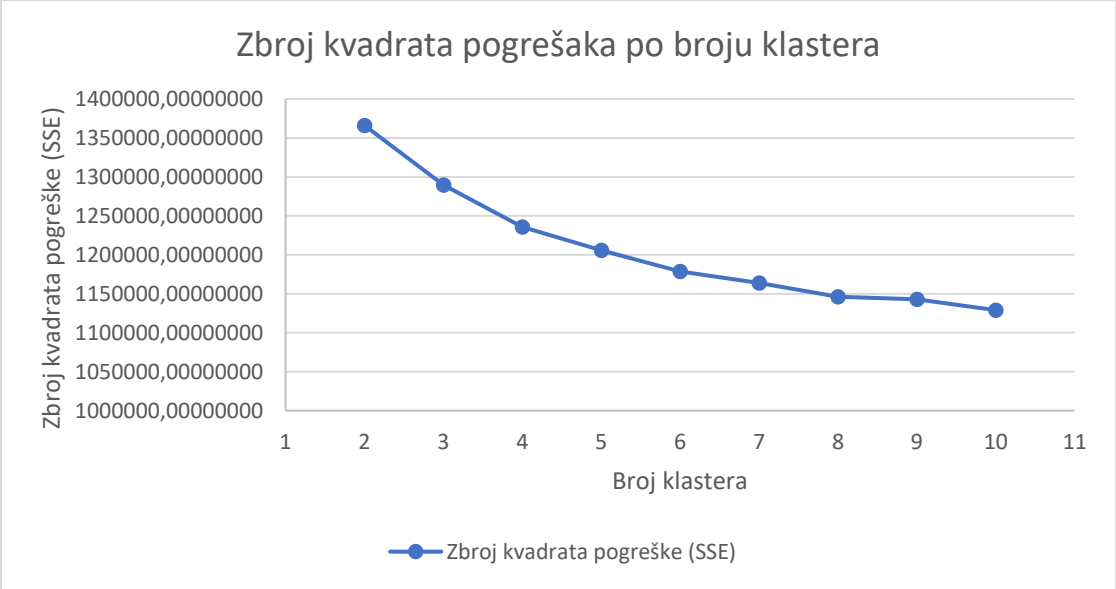
Zatim se odabire panel Cluster u kojem se provodi klaster analiza. Cilj klaster analize je grupiranje osoba u klastere prema socio-demografskim čimbenicima i ostalim zajedničkim čimbenicima kako bi se identificirali uzroci i povezanosti koji doprinose razvoju dijabetesa. Za izradu klaster analize je korištena K-means metoda odnosno algoritam k-srednjih vrijednosti.

Broj klastera je odabran pomoću metode lakta (eng. *Elbow method*). Metoda lakta je tehnika koja se koristi u strojnom učenju, a pogotovo u klasterizaciji podataka. Primjenjuje se za

određivanje optimalnog broja klastera prilikom provođenja klaster analize. Cilj navedene metode je pronaći točku odnosno „lakat“ na grafikonu koji označava najbolji broj klastera za segmentaciju podataka. Ovaj postupak pomaže vizualno identificirati točku u kojoj dodavanje dodatnih klastera prestaje značajno smanjivati sumu kvadrata pogreške (SSE), čime se određuje optimalan broj klastera za daljnju analizu.

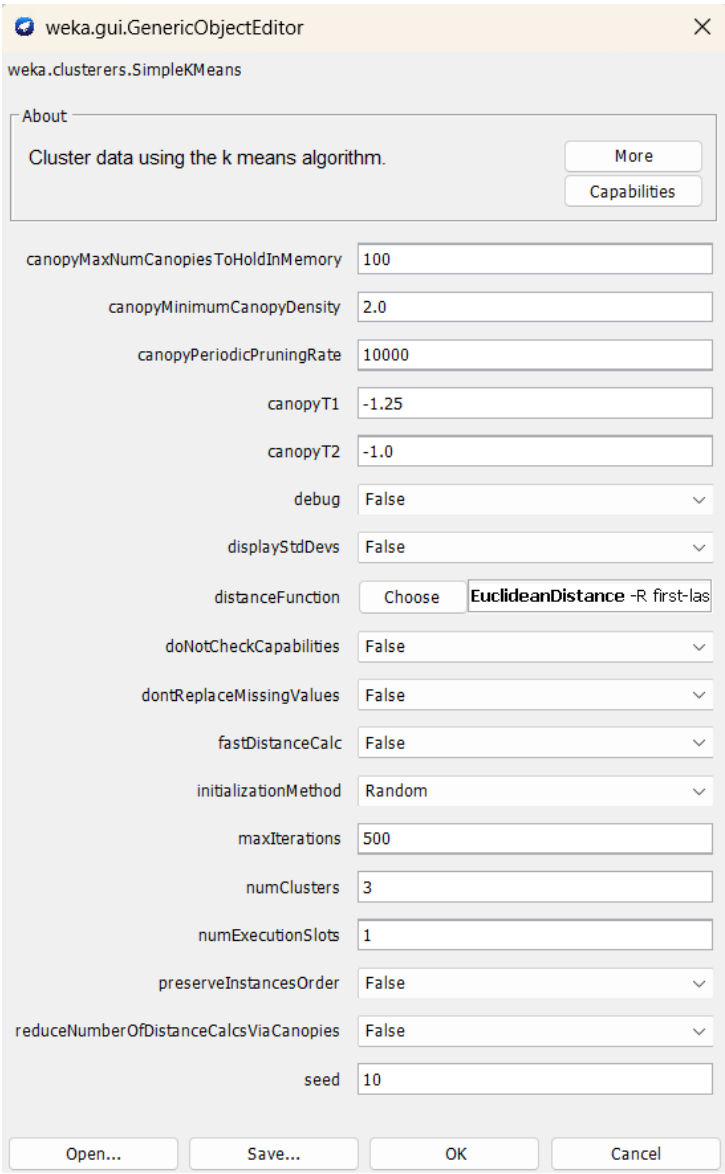
Kako bi se primijenila metoda lakta, provedena je analiza zbroja kvadrata pogreške za različite brojeve klastera. Na temelju izračunatih zbrojeva kvadrata pogrešaka za različite brojeve klastera, kreiran je grafikon koji prikazuje kako se zbroj kvadrata pogreške mijenja s brojem klastera. Na grafikonu je na X-osi prikazan broj klastera, a na Y-osi zbroj kvadrata pogreške. Kako bi se odabrao optimalan broj klastera za provedbu klaster analize, potrebno je identificirati točku gdje dolazi do značajne promjene u nagibu krivulje, nakon čega dodavanje dodatnih klastera ne rezultira značajnim smanjenjem zbroja kvadrata pogreške.

Na prikazanom grafikonu, nakon trećeg klastera pad zbroja kvadrata pogreške postaje manje izražen. Prema metodi lakta, optimalan broj klastera je 3.



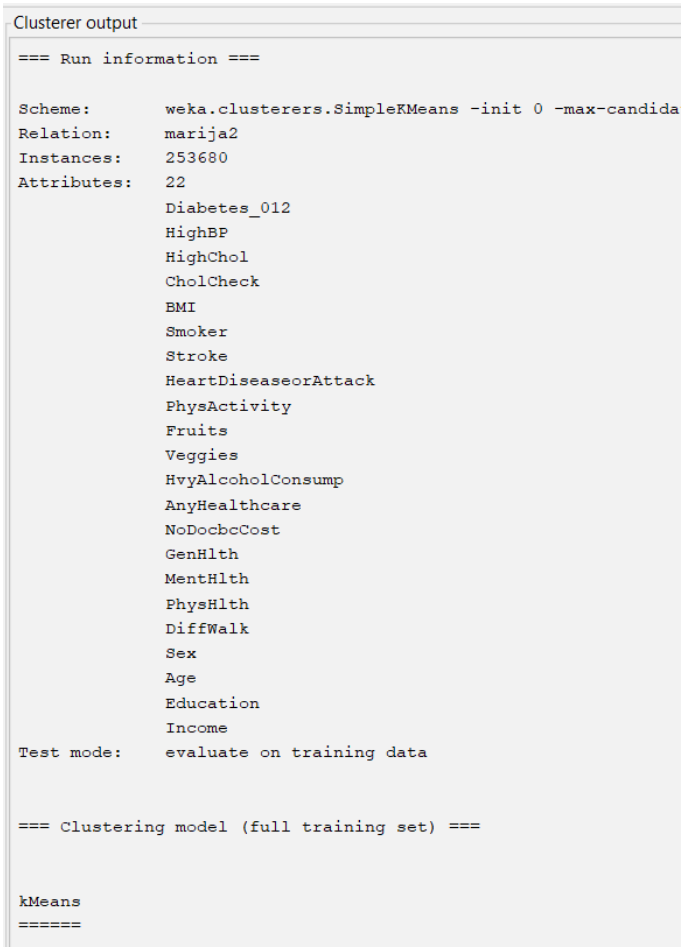
Slika 24 Grafički prikaz zbroja kvadrata pogrešaka po broju klastera

Klasni atribut je Diabetes_012 koji označava ima li osoba dijabetes, predijabetes ili nema dijabetes. Ostali atributi se koriste za grupiranje i analizu podataka u svrhu identifikacije ključnih čimbenika koji utječu na dijabetes i pronalasku sličnih grupa ljudi prema njihovim socio-demografskim čimbenicima, životnim navikama i fizičkom stanju. Svaki klaster će predstavljati grupu ljudi sa sličnim karakteristikama i načinom života čime će se pokušati otkriti specifični čimbenici koji utječu na razvoj bolesti.



Slika 25 Odabir Simple K Means metode

Nakon što je odabrana Simple K Means metoda i broj klastera pokreće se analiza, a rezultati su vidljivi u okviru Clusterer output.



Slika 26 Informacije o setu podataka

Slika 26 prikazuje osnovne podatke o korištenom setu podataka. U izradi klasifikacijskog modela korišten je algoritam K-srednjih vrijednosti. Na slici se vidi kako je korišteno 253680 primjera te da set podataka sadrži 22 atributa. Prvi atribut Diabetes_012 je klasni atribut.

Klasifikacijski model izrađen na temelju full training seta podataka pretvoren je u tablični prikaz. Model je izgradio ukupno 3 klastera koji su označeni brojevima 0, 1 i 2. Svaki klaster ima određenu distribuciju instanci iz skupa podataka. Klaster 0 sadrži 46219 instanci, klaster 1 ima 75772 instanci i klaster 2 ima 131689 instanci.

Tablica 2 sadrži prikaz klastera.

Klaster	0	1	2
Diabetes_012	no_diabetes	no_diabetes	no_diabetes
HighBP	high_BP	no_high_BP	no_high_BP
HighChol	high_cholesterol	no_high_cholesterol	no_high_cholesterol
CholCheck	yes_cholesterol_check_in_5_years	yes_cholesterol_check_in_5_years	yes_cholesterol_check_in_5_years
BMI	31.1081	28.5579	27.3247
Smoker	yes	yes	no
Stroke	no	no	no
HeartDiseaseorAttack	no	no	no
PhysActivity	no	yes	yes
Fruits	no	yes	yes
Veggies	yes	yes	yes
HvyAlcholConsump	no	no	no
AnyHealthcare	yes	yes	yes
NoDocbcCost	no	no	no
GenHlth	fair	good	very_good
MentHlth	7.3042	2.9354	1.8825
PhysHlth	12.2588	3.6299	1.7807
DiffWalk	yes	no	no
Sex	female	male	female
Age	60_to_64	65_to_69	60_to_64
Education	some_college_or_technical_school	college_graduate	college_graduate
Income	Less_than_$50000	$75000_or_more	$75000_or_more

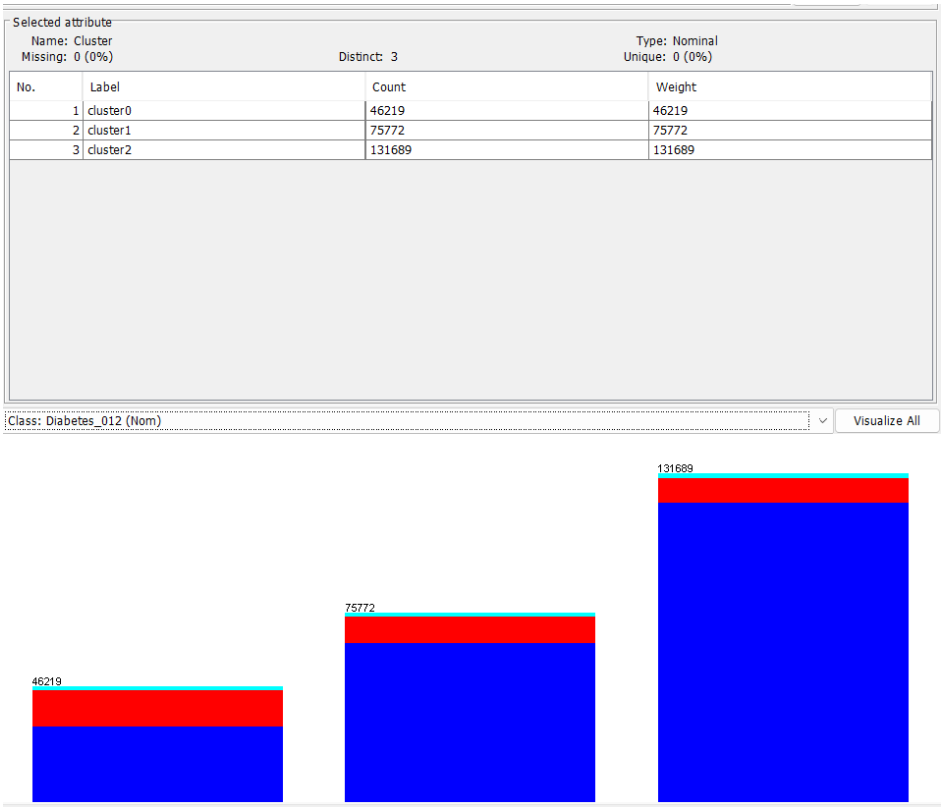
Tablica 2 Tablični prikaz klastera

Izvor: Rad autora

Navedeni klasteri predstavljaju grupiranje sličnih uzoraka ili karakteristika unutar skupa podataka stoga instance u istom klasteru mogu dijeliti slične demografske čimbenike, fizičko stanje i životne navike koje su uključene u skupu podataka.

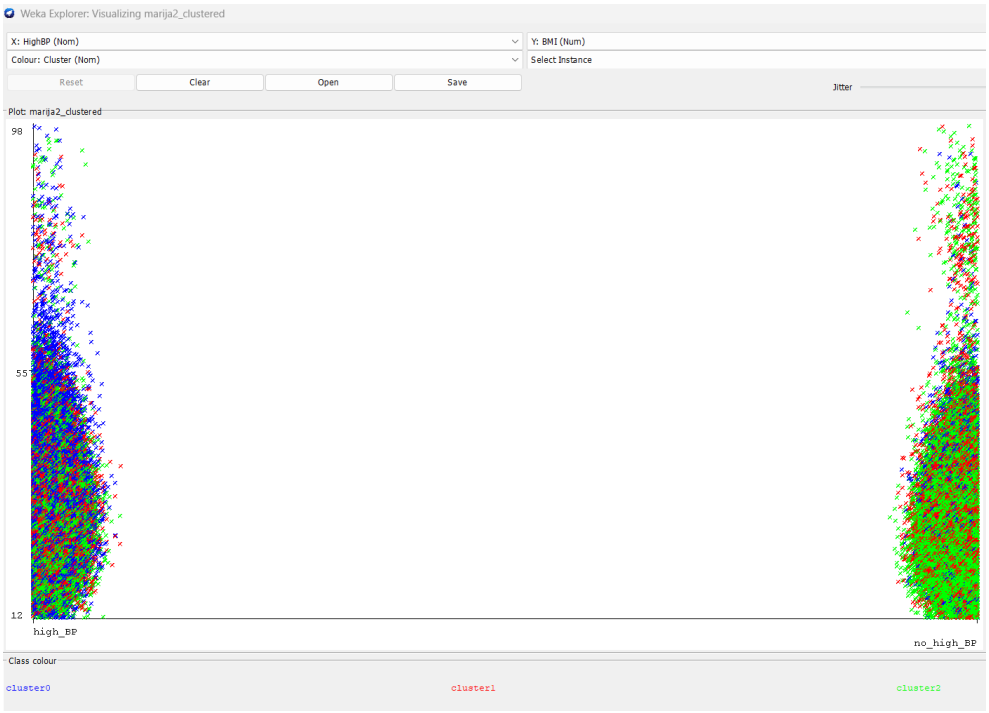
Svaki klaster ima svoj centroid koji predstavlja prosječne vrijednosti svih atributa za instance unutar tog klastera. K-means analiza je provedena s ukupno 6 iteracija, a suma kvadrata udaljenosti unutar klastera (eng. *Within cluster sum of squared errors*) iznosi 1289540.1189297317. Ova mjera pokazuje koliko su točke unutar istog klastera udaljene jedna od druge odnosno koliko su točke unutar klastera blizu svojem središtu ili centroidu.

Prikaz klastera i atributa omogućuje razumijevanje distribucije instanci unutar svakog klastera. Dobiveni klasteri su pohranjeni u Weki u .arff formatu te su ponovno učitani u Weku čime je nastao novi atribut „Cluster“. Dobiveni atribut pokazuje pripadnost instanci odnosno primjera svakom klasteru. Na slici 27 su prikazani klasteri u panelu Preprocess, a vidljivo je kako klasteru 0 odnosno prvom klasteru pripada 46.219 primjera, klasteru 1 odnosno drugom klasteru 75.772 primjera te klasteru 2 odnosno drugom klasteru 131.689 primjera. Iz prikazanog grafa je vidljivo kako najveći broj oboljelih pripada prvom klasteru.



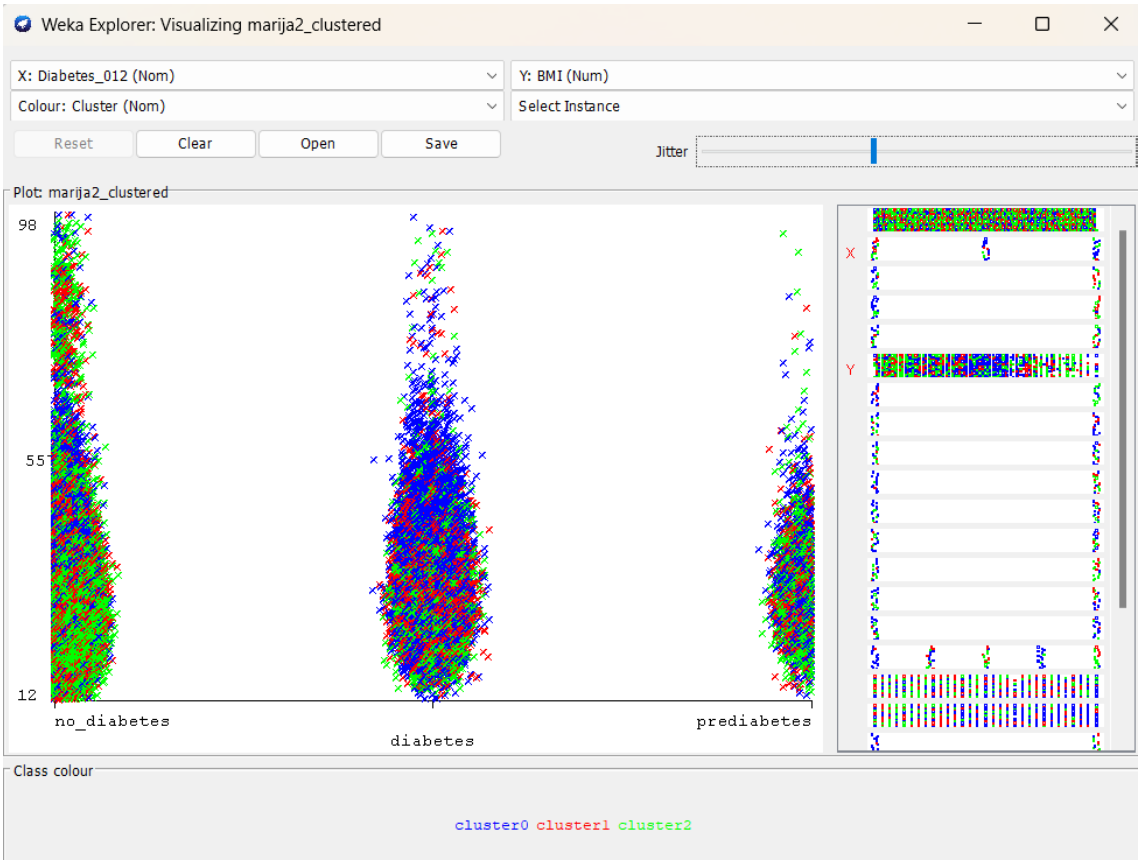
Slika 27 Prikaz klastera u Preprocess panelu

Pripadnost vrijednosti klasterima vizualizirana je opcijom Visualize cluster segments. Na slici 28 prikazan je graf izrađen na temelju atributa. Navedeni graf prikazuje rasprostranjenost određenih klastera pomoću atributa HighBP (prikazan na X-osi) i atributa BMI (prikazan na Y-osi). Klasteri su označeni različitim bojama: plava (Cluster 0), crvena (Cluster 1) i zelena (Cluster 2). Dobiveni graf prikazuje kako su različiti klasteri raspoređeni po BMI-ju i prisutnosti visokog krvnog tlaka. Pojedinci s visokim krvnim tlakom većinom su grupirani u klaster 0 dok su oni s normalnim krvnim tlakom grupirani u klastere 1 i 2.



Slika 28 Visualize cluster segments

Na slici 29 su vizualizirani rezultati klasterizacije pomoću atributa Diabetes_012 (prikazan na X-osi) i BMI-ja (prikazan na Y-osi). Iz grafa se može zaključiti kako osobe bez dijabetesa imaju BMI vrijednosti koje su koncentrirane u nižem rasponu, dok osobe s dijabetesom i predijabetesom imaju širi raspon BMI vrijednosti što sugerira da postoji veća varijabilnost u tjelesnoj masi među dijabetičarima.



Slika 29 Visualize cluster segments

4.3. Rasprava i prijedlozi

Provedena analiza omogućava razumijevanje grupiranja sudionika istraživanja prema sličnim čimbenicima. Ova analiza može poslužiti za stjecanje boljeg razumijevanje čimbenika dijabetesa čime se mogu razviti strategije za prevenciju i liječenje bolesti. Svaki klaster će predstavljati specifičnu grupu ljudi s karakterističnim čimbenicima, navikama i ponašanjima.

Vidljivo je kako su socio-demografski i drugi čimbenici grupirani u tri klastera (0-2).

Prvi klaster odnosno klaster 0 karakteriziraju osobe koje nemaju dijabetes, imaju visoki krvni tlak i visok kolesterol. Pripadnici su u posljednjih pet godina obavili provjeru kolesterola. Indeks tjelesne mase je uredan. Ispitanici u ovome klasteru su većinom pušači, nisu imali moždani udar niti koronarnu bolest ili infarkt miokarda. Ispitanici se u posljednjih 30 dana nisu bavili tjelesnom aktivnošću te nisu konzumirali voće. Sudionici većinom konzumiraju povrće. Sudionici imaju barem neki oblik zdravstvene zaštite. Sudionici nemaju naviku teške konzumacije alkohola. Opće stanje zdravlja je uredno. Tijekom proteklih 30 dana mentalno zdravlje nije bilo dobro 7,3042 dana, a fizičko zdravlje nije bilo dobro 12,2588 dana. Pripadnici ovog klastera imaju poteškoća s hodanjem ili penjanjem stepenicama. Osobe unutar ovog

klastera su pretežno ženskog spola i dobi od 60 do 64 godine. Većina sudionika unutar klastera ima završen neki fakultet ili tehničku školu. Razina dohotka je manja od 50 000 $.

Drugi klaster odnosno klaster 1 karakteriziraju osobe koje nemaju dijabetes, nemaju visoki krvni tlak, nemaju visok kolesterol te osobe koje su u posljednjih pet godina obavile provjeru kolesterola. Indeks tjelesne mase je uredan. Ispitanici su većinom pušači. Osobe koje su sudjelovale u istraživanju većinom nisu imali moždani udar niti koronarnu bolest ili infarkt miokarda. Osobe koje pripadaju ovom klasteru su se u posljednjih 30 dana bavile tjelesnom aktivnošću te su konzumirale voće i povrće. Sudionici nemaju naviku teške konzumacije alkohola. Opće stanje zdravlja je dobro. Tijekom proteklih 30 dana mentalno zdravlje nije bilo dobro 2,9354 dana, a fizičko zdravlje nije bilo dobro 3,6299 dana. Osobe koje pripadaju ovom klasteru nemaju poteškoće s hodanjem ili penjanjem stepenicama. Pripadnici ovog klastera su većinom muškog spola i u dobi od 65 do 69 godina. Većina sudionika je diplomirala na fakultetu. Razina dohotka je 75 000 $ ili više.

Treći klaster odnosno klaster 2 karakteriziraju osobe koje nemaju dijabetes te koje nemaju visoki krvni tlak i visoki kolesterol te osobe koje su u posljednjih pet godina obavile provjeru kolesterola. Indeks tjelesne mase je uredan, manji u odnosu na ostale klastere. Osobe u ovom klasteru nisu pušači. Osobe koje su sudjelovale u istraživanju većinom nisu imali moždani udar niti koronarnu bolest ili infarkt miokarda. Pripadnici ovog klastera su se posljednjih 30 dana bavili tjelesnom aktivnošću te većinom konzumiraju voće i povrće. Sudionici nemaju naviku teške konzumacije alkohola. Opće stanje zdravlja je vrlo dobro. Stanje mentalnog zdravlja tijekom posljednjih 30 dana nije bilo dobro 1,8825 dana, a fizičkog zdravlja 1,7807 dana. Pripadnici klastera nemaju poteškoća s hodanjem ili penjanjem stepenicama. Pripadnici su većinom ženskog spola u dobi od 60 do 64 godina. Većina sudionika je diplomirala na fakultetu, a razina dohotka je 75 000 ili više.

Provedena analiza daje dublji uvid u karakteristike različitih skupina sudionika istraživanja te u njihove veze s rizikom od bolesti dijabetesa. Iz analize se može zaključiti kako svaki klaster ima specifične poveznice s rizikom od bolesti i općenitim zdravstvenim navikama. Vidljivo je kako prvi klaster čine osobe koje imaju visoki krvni tlak i kolesterol, koje nisu fizički aktivne. S druge strane, pripadnici drugog i trećeg klastera su osobe s nižim rizikom od nastanka bolesti, aktivnije osobe i osobe s zdravijim prehrambenim navikama. Također, vidljive su i socio-demografske razlike između klastera pa tako prvi klaster većinom čine žene starije dobi s nižim prihodima i nižim stupnjem obrazovanja. Drugom i trećem klasteru većinom pripadaju muškarci s višim prihodima i višim stupnjem obrazovanja.

Drugi i treći klaster ukazuju na važnost tjelesne aktivnosti i zdravih prehrambenih navika i na njihovu povezanost s manjim rizikom od bolesti. Vidljivo je kako su ovi klasteri aktivniji i kako konzumiraju više voća i povrća što je, kako je spomenuto, vrlo značajno u prevenciji dijabetesa.

Dodatno, na temelju rezultata provedene analize, može se zaključiti kako se kod pojedinih skupina može pojaviti potreba za dodatnim strategijama liječenja ili prevencije bolesti. Po rezultatima atributa koji se odnosi na dane lošijeg mentalnog zdravlja, prvom klasteru bi mogli biti korisni programi za upravljanje stresom i ostalim psihičkim stanjima dok bi se za druge klastere mogle razviti programi koji su usmjereni na poboljšanje općeg zdravlja.

Nadalje, preporučuje se razvoj programa za kontrolu visokog krvnog tlaka i kolesterola, posebno za skupinu s visokim BMI-jem i visokim prihodima te edukacijske kampanje o važnosti pravilne prehrane, tjelesne aktivnosti i prestanka pušenja za mlađe osobe s dijabetesom i nižim prihodima. Potrebno je potaknuti fizičku aktivnost kod svih dobnih skupina, posebno među onima s višim prihodima te promovirati redovite zdravstvene preglede kako bi se spriječili mogući zdravstveni problemi. Ulaganje u obrazovanje i svijest o zdravlju, posebno među mlađim skupinama, može značajno doprinijeti boljem zdravstvenom stanju populacije.

Provedena analiza omogućava identificiranje specifičnih potreba svakog klastera čime je moguće osmisliti prevencijske programe usmjerene na takve potrebe. Rezultati analize daju osnovu za daljnja istraživanja kako bi se relevantni čimbenici bolje razumjeli i kako bi se razvile efikasnije metode prevencije i liječenja bolesti.

5. ZAKLJUČAK

Dijabetes je kronični metabolički poremećaj karakteriziran povišenim razinama glukoze u krvi koji može dovesti do ozbiljnih zdravstvenih komplikacija. Bolest se može podijeliti na nekoliko različitih vrsti od kojih je najčešći dijabetes tipa 2. U današnje vrijeme pojava dijabetesa ubrzano raste zbog različitih čimbenika, kao što su sjedilački način života, nezdrave prehrambene navike i pretilost, što predstavlja veliki izazov za zdravstvene sustave diljem svijeta.

U radu je predstavljen i opisan proces otkrivanja znanja iz baza podataka te je opisana važnost primjene otkrivanja znanja kod bolesti dijabetesa, kao i u zdravstvu općenito. Zaključeno je kako se učinkovitim korištenjem rudarenja podataka u zdravstvenom sustavu lakše donose ispravne odluke temeljene na analizi velikih skupova podataka, optimiziraju se resursi te se pruža kvalitetnija skrb o pacijentima što posljedično poboljšava ishode za pacijente i cijeli sustav.

Provedena je klaster analiza na temelju očišćenog i konsolidiranog skupa podataka „Diabetes Health Indicators Dataset" s ciljem grupiranja sudionika u slične klasteru na temelju njihovih socio-demografskih čimbenika, životnih navika i fizičkog stanja. Analiza je dala dublji uvid u karakteristike različitih klastera te u veze povezane s rizikom od nastanka bolesti. Može se zaključiti kako svaki klaster ima specifične poveznice s rizikom od bolesti i općenitim zdravstvenim navikama.

Rezultati analize pokazuju kako prvi klaster čine osobe s visokim krvnim tlakom i kolesterolom te osobe koje nisu fizičke aktivne dok su pripadnici drugog i trećeg klastera osobe s nižim rizikom od razvoja bolesti, aktivnije osobe i osobe sa zdravijim prehrambenim navikama. Isto tako, vidljive su i razlike u socio-demografskim čimbenicima. Prvi klaster većinski čine ženske osobe starije dobe s nižim prihodima i nižim stupnjem obrazovanja dok drugom i trećem klasteru pripadaju muške osobe s većim prihodima i višim stupnjem obrazovanja.

Dodatno, na temelju rezultata analize, može se zaključiti kako se kod nekih skupina pojavljuje potreba za dodatnim strategijama liječenja ili prevencije bolesti. Analiza omogućava identificiranje specifičnih potreba svakog klastera na temelju kojih se mogu osmisliti prevencijski programi usmjereni na takve potrebe. Rezultati daju temelj za daljnja istraživanja kako bi se svi relevantni čimbenici mogli bolje razumjeti i kako bi se razvile učinkovite metode prevencije i liječenja.

POPIS LITERATURE

1. Al Yousef, M. Z., Yasky, A. F., Al Shammari, R., & Ferwana, M. S. (2022). Early prediction of diabetes by applying data mining techniques: A retrospective cohort study. *Medicine*, *101*(29), e29588.
2. Centers for Disease Control and Prevention (2023.) What is Diabetes?, preuzeto 17. travanj 2024. s https://www.cdc.gov/diabetes/basics/diabetes.html
3. Chait, A., & Den Hartigh, L. J. (2020). Adipose tissue distribution, inflammation and its metabolic consequences, including diabetes and cardiovascular disease. Frontiers in cardiovascular medicine, 7, 522637.
4. Clark, N. G., Fox, K. M., Grandy, S., & SHIELD Study Group. (2007). Symptoms of diabetes and their association with the risk and presence of diabetes: findings from the Study to Help Improve Early evaluation and management of risk factors Leading to Diabetes (SHIELD). Diabetes Care, 30(11), 2868-2873.
5. Diabetes UK. (n.d.). Differences between type 1 and type 2 diabetes, preuzeto 16. travnja 2024. s https://www.diabetes.org.uk/diabetes-the-basics/differences-between-type-1-and-type-2-diabetes
6. Haire-Joshu, D., Glasgow, R., Tibbs, T. (1999.) Smoking and Diabetes. Diabetes Care. Volume 22 (11)
7. Han, J., Kamber, M. & Pei, J. (2012) *Data mining: Concepts and techniques.* Third edition. Waltham: Morgan Kaufmann Publishers.
8. Harvard T.H. Chan (2021.) Simple steps to preventing diabetes, preuzeto 17. travnja 2024. s https://www.hsph.harvard.edu/nutritionsource/disease-prevention/diabetes-prevention/preventing-diabetes-full-story/
9. Jain, N., & Srivastava, V. (2013). Data mining techniques: a survey paper. IJRET: International Journal of Research in Engineering and Technology, 2(11), 2319-1163.
10. Kaggle (2021.) Diabetes Health Indicators Dataset [Data file], preuzeto s https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/dana
11. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, *15*, 104-116.
12. Kyrou, I., Tsigos, C., Mavrogianni, C., Cardon, G., Van Stappen, V., Latomme, J., ... & Manios, Y. (2020). Sociodemographic and lifestyle-related risk factors for identifying

vulnerable groups for type 2 diabetes: a narrative review with emphasis on data from Europe. *BMC endocrine disorders*, *20*, 1-13.

13. Lutkevich, B. (2023.) TechTarget. Association rules, preuzeto 25. travnja 2024. s https://www.techtarget.com/searchbusinessanalytics/definition/association-rules-in-data-mining
14. Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. In *Data mining and knowledge discovery in real life applications*. IntechOpen.
15. Pejić Bach, M. (2005.) Rudarenje podataka u bankarstvu. *Zbornik ekonomskog fakulteta u Zagrebu*, *3*(1), 181-193.
16. Pejić Bach, M. (2007.) Otkrivanje znanja iz baza podataka.
17. Pejić Bach, M., & Kerep, I. (2011). Weka–alat za otkrivanje znanja iz baza podataka.
18. Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. *Measurement: Sensors*, *25*, 100605.
19. Roglic, G. (2016). WHO Global report on diabetes: A summary. *International Journal of Noncommunicable Diseases*, *1*(1), 3-8.
20. Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.
21. Diabetes Prevention Program (DPP) Research Group. (2002). The Diabetes Prevention Program (DPP) description of lifestyle intervention. Diabetes care, 25(12), 2165-2171.
22. Vukić, I., Pravdić, D. (2020.) Važnost pravilne prehrane osoba oboljelih od šećerne bolesti. Fakultet zdravstvenih studija. Sveučilište u Mostaru.
23. Witten, I., Frank, E., Hall, M. (2016.) Data Mining: Practical Machine Learning Tools and Techniques. Fourth Edition. Elsevier.
24. Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. Applied artificial intelligence, 17(5-6), 375-381.

POPIS SLIKA

POPIS TABLICA

ŽIVOTOPIS

MARIJA BURIĆ

marijaburic7@gmail.com

14.04.1999. | Anđela Nuića 28, 10040 Zagreb | +385977240561

OBRAZOVANJE

Ekonomski fakultet, Sveučilište u Zagrebu | 2018. - trenutno
Integrirani preddiplomski i diplomski sveučilišni studij Poslovne ekonomije
smjer: Menadžerska informatika

Gornjogradska gimnazija | 2014. - 2018.

RADNO ISKUSTVO

Ernst & Young, Poslovno savjetovanje

IT revizija | listopad 2023 - trenutno

- Pregled, testiranje i dokumentiranje IT generalnih kontrola iz područja upravljanja korisničkim računima, upravljanja promjenama i upravljanja IT operacijama
- Testiranje aplikativnih kontrola
- Evaluacija i testiranje kontrola prema EY metodologiji, lokalnoj regulativi te svjetskim praksama
- Pisanje regulatornog izvješća

OTP Leasing, Direkcija IT i digitalne transformacije poslovanja

Rad u IT odjelu | svibanj 2023 - rujan 2023

- Administrativni poslovi
- Odobravanje računa
- Rješavanje Help Desk zahtjeva
- Pisanje dokumentacije i pravilnika u svrhu zatvaranja preporuka IT revizije

Zagrebačka banka, Korporativno bankarstvo

Administrativni poslovi | srpanj 2022 - rujan 2022

- Pomoć u administriranju dokumentacije
- Rješavanje upita i reklamacija klijenata
- Podrška timu kod obavljanja poslova otvaranja računa, ugovaranja novih usluga, izmjene postojećih usluga
- Komunikacija i suradnja s ostalim dijelovima Banke s ciljem pružanja kompletne usluge klijentima

Bonbon

Agent korisničke podrške | listopad 2021 - lipanj 2022

- Zaprimanje i rješavanje zahtjeva, upita i prigovora korisnika
- Prodaja usluga
- Administracija poziva

VJEŠTINE

- komunikacijske vještine
- organizacijske vještine
- MS Office
- C1 Engleski jezik
- A1 Talijanski jezik
- Osnovno poznavanje - Bizagi Process Modeler, Camunda, SQL, C#, Weka