Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko

Seminarska naloga

VITKO UPRAVLJANJE

David Dugar

Maribor, januar 2025

Vsebina

UVOD

Vitko upravljanje, znano tudi kot Lean Management, je metoda, osredotočena na izboljšanje procesov z odpravljanjem nepotrebnih korakov ter zagotavljanjem stabilnosti in učinkovitosti. Temelj tega pristopa je t.i. "pull model", ki omogoča izvajanje procesov na podlagi dejanskih potreb. Metodologija se pogosto uporablja v avtomobilski industrijski, zdravstveni, storitveni indrustriji in številnih drugih sektorjih. Organizacije, ki sledijo filozofiji stalnih izboljšav, uporabljajo vitko upravljanje za uvajanje manjših sprememb, s katerimi dosežejo optimizacijo specifičnih procesov. Pogosto je vitko upravljanje združeno z metodo Six Sigma, ki se osredotoča na zmanjševanje variabilnosti v procesih in s tem povečevanje dobičkonosnosti podjetja. Six Sigma temelji na podatkovni analitiki in uporabi statističnih metod, pri čemer je najbolj prepoznaven okvir DMAIC (Definiraj, Meri, Analiziraj, Izboljšaj, Kontroliraj). Ta pristop omogoča strukturirano in sistematično izboljševanje procesov ter odpravljanje neučinkovitosti.

Diabetis je resna kronična bolezen, pri kateri pacient/ka ne more efektivno regulirati nivoje glukoze v lastni krvi. Bolezen lahko vodi do znižane življenske kvalitete ali pa celo do krajšega življenskega obdobja. Po tem, ko se hrana v želodcu prebavi se razgradi na sladkorje, ki se potem spustijo v krvni obtok. Povišanje tega krvnega sladkorja pa telo preko trebušne slinavke zazna in v krvni obtok pošlje encim imenovan inzulin. Inzulin pomaga pri reguliranju nivoja sladkorja v krvi, tako da telesnim celicam omogoči porabo tega sladkorja za energijo/delovanje. Diabetis pa pomeni, da pacient oz. pacientka ne sprošča dovolj inzulina v krvni obtok ali pa ne more sproščenega inzulina efektino uporabiti za njegove namene. Poškodbe oz. komplikacije kot srčne bolezni, izguba vida, amputacija spodnjih okončin in bolezni ledvic so povezane z kroničnim ohranjanjem visokega nivoja sladkorja v krvi za diabetike. Zaenkrat še ne obstaja nekega zdravila, ki bi direktno ozdravela diabetis. Obstajajo pa načini oz. strategije, ki reducirajo škodo ali pa celo nastanek diabetisa. Nekateri izmed teh strategij so izguba telesna mase, zdrava prehrana, telesna aktivnost in nudenje zdravstvene nege. Zgodnja diagnoza diabetisa lahko vodi do signifikantnih sprememb življenskega načina, ter učinkovitejšega zdravljenja pacienta.

CDC (The Centers for Disease Control) ocenjuje, da se 1 od 5 diabetikov in približno 8 od 10 preddiabetikov ne zaveda svojega tveganja. Čeprav obstajajo različne vrste sladkorne bolezni, je sladkorna bolezen tipa II najpogostejša oblika, njena razširjenost pa se razlikuje glede na starost, izobrazbo, dohodek, lokacijo in druge družbene dejavnike zdravja. Velik del bremena bolezni pade tudi na tiste z nižjim socialno-ekonomskim statusom. Sladkorna bolezen predstavlja tudi veliko breme za gospodarstvo, saj stroški diagnosticirane sladkorne bolezni znašajo približno 327 milijard dolarjev, skupni stroški z nediagnosticirano sladkorno boleznijo in prediabetesom pa se približujejo 400 milijardam dolarjev letno.

Koncepte vitkega upravljanja smo uporabili na realnih podatkih. Za podatkovno bazo smo si izbrali BRFSS (The Behavioral Risk Factor Surveillance System), ki je letna telefonska anketa o zdravju, zbrana pod okriljem CDC (The Centers for Disease Control). Ta anketa se izvaja že od leta 1984 in vsako leto vključuje odgovore več kot 400.000 Američanov. Namenjena je zbiranju podatkov o tveganem vedenju, povezanem z zdravjem, kroničnih zdravstvenih stanjih ter uporabi preventivnih storitev. Za namen projekta smo uporabili CSV nabor podatkov BRFSS, ki je dostopen na platformi Kaggle za leto 2015 . Ta baza vključuje odgovore ameriških anketirancev in je usmernjena na pridobivanje čim več podatkov povezanih s sladkorno boleznijo.

Glavni cilj te naloge je, da s pomočjo pristopov vitkega upravljanja optimiziramo dva izbrana problema.

OPIS PODATKOV IN PROBLEMA

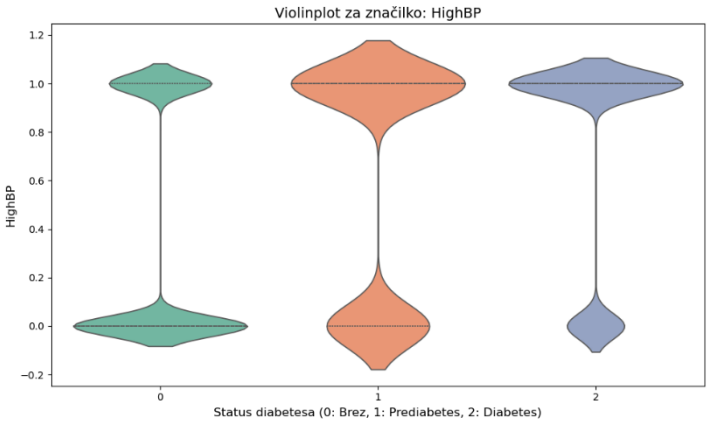
V podatkovni bazi je 22 spremenljivk, ki vsebujejo informacije o različnih dejavnikih na sladkorno bolezen.

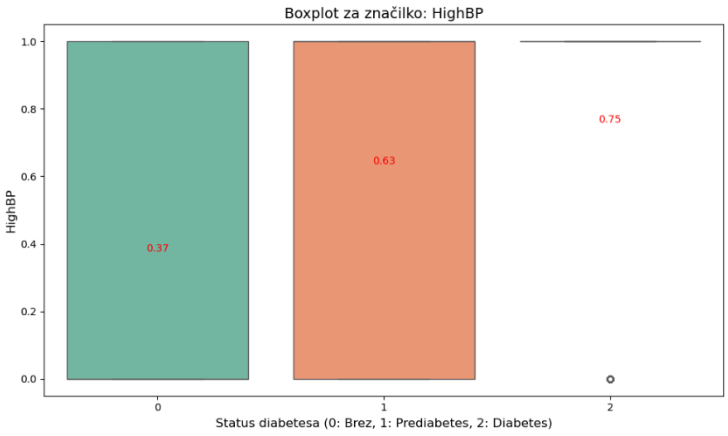
Spremenljivka Diabetes_012

Ta spremenljivka nam pove ali oseba ima sladkorno bolezen.
Spremenljivka je nominalna in številčno kodirana.
0 - nediabetik
1 - prediabetik
2 - diabetik

Spremenljivka HighBP

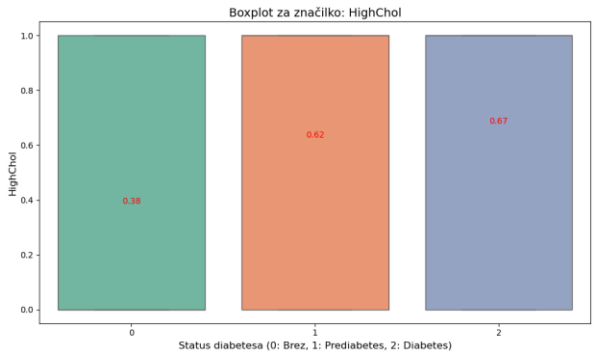
Odrasli, ki jim je zdravnik, medicinska sestra ali drug zdravstveni delavec povedal, da imajo visok krvni tlak.
Spremenljivka je nominalna in številčno kodirana.
0 - ne
1 – da

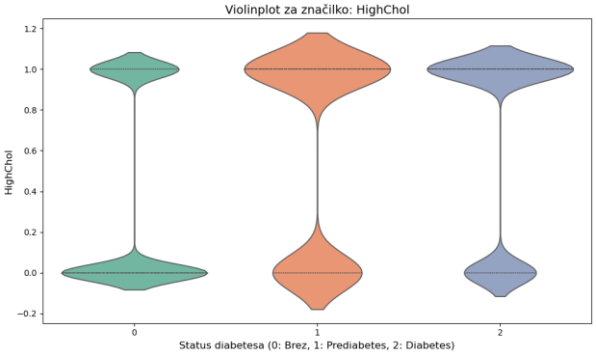




Spremenljivka HighChol

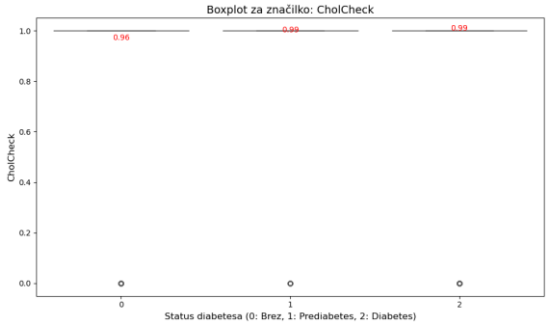
Odrasli, ki so jim izmerili holesterol in jim je to povedal zdravnik, medicinska sestra ali drug zdravstveni delavec, da je bilo visoko.
Spremenljivka je nominalna in številčno kodirana.
0 – ni bilo visoko
1 – je bilo visoko

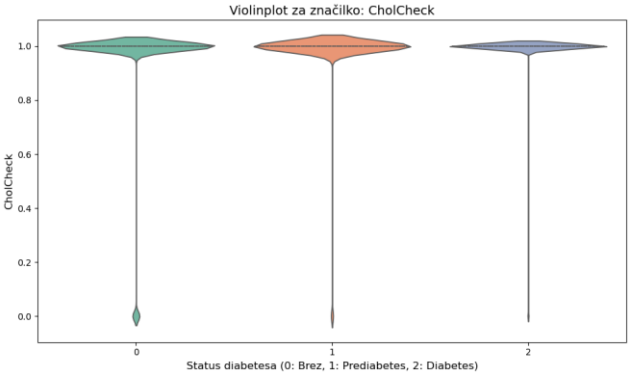




Spremenljivka CholCheck

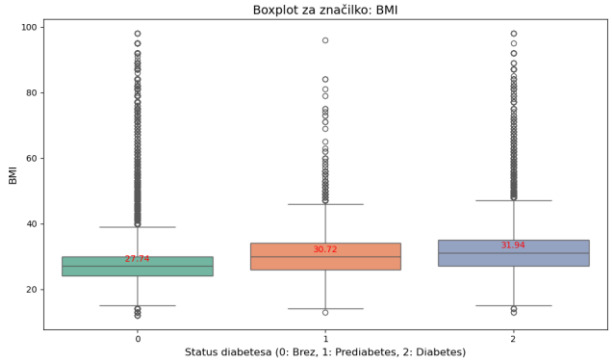
Ali je oseba imela preverjanje holesterola v zadnjih petih letih.
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da

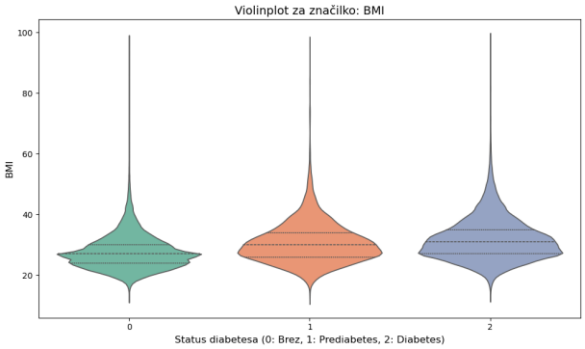




Spremenljivka BMI (ITM)

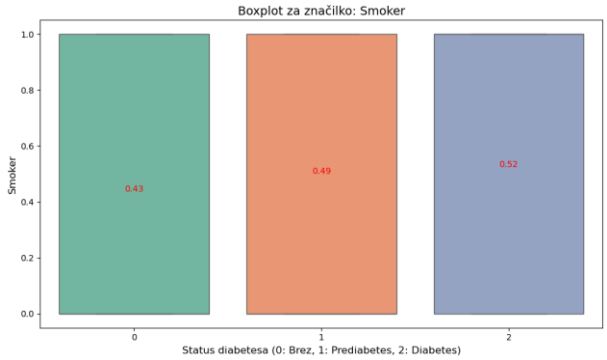
Pove nam indeks telesna mase anketirane osebe.
Spremenljivka je kvantitativna.

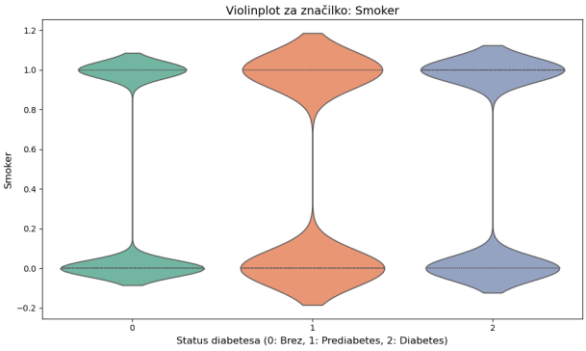




Spremenljivka Smoker

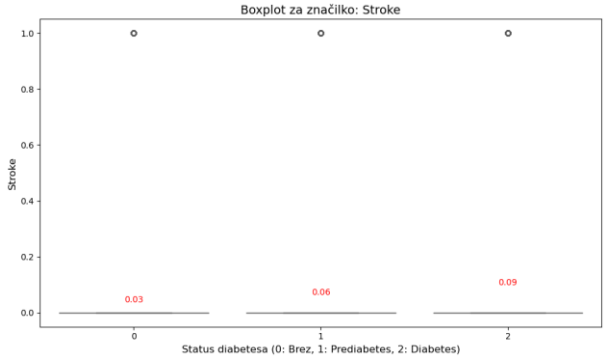
Ali je oseba v svojem življenju skadila vsaj 100 cigaret.
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da

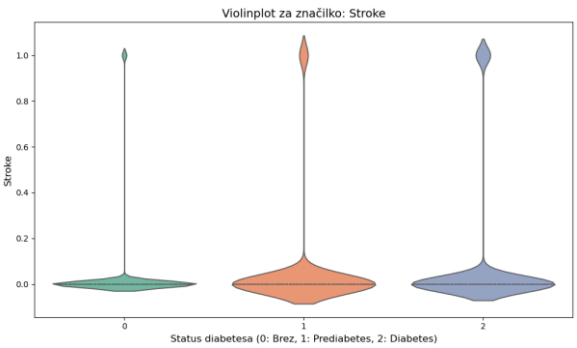




Spremenljivka Stroke

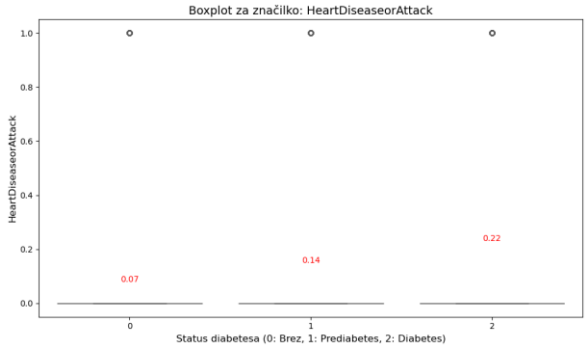
Ali je oseba kdajkoli doživela kap.
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da

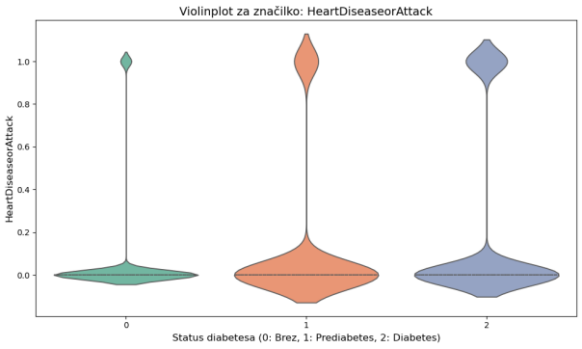




Spremenljivka HeartDiseaseorAttack

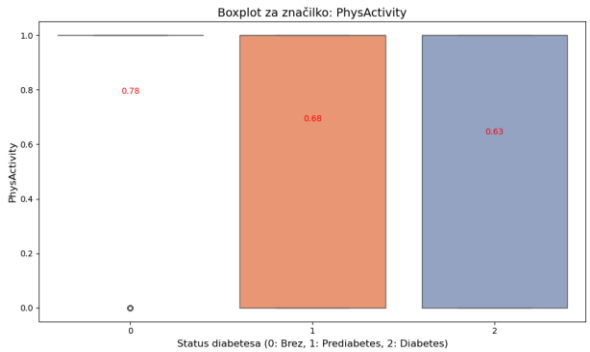
Ali oseba ima CHD (Coronary heart disease) ali MI (Myocardial Infarction)
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da

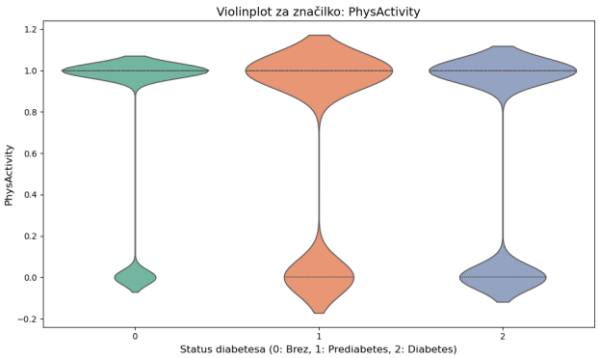




Spremenljivka PhysActivity

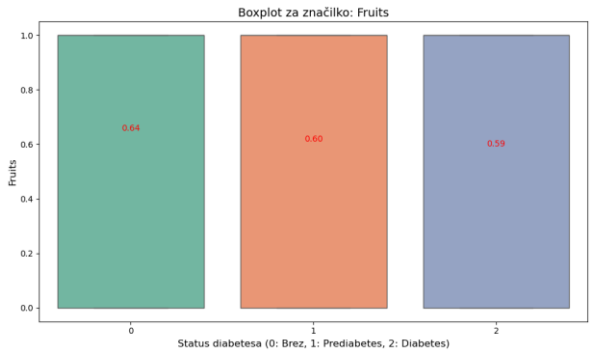
Ali je oseba bila fizično aktivna v zadnjih 30 dneh, razen v službi
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da

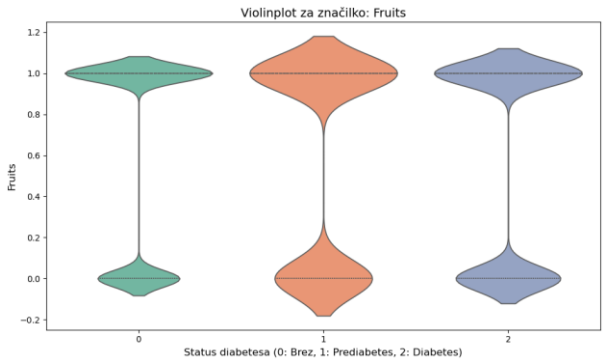




Spremenljivka Fruits

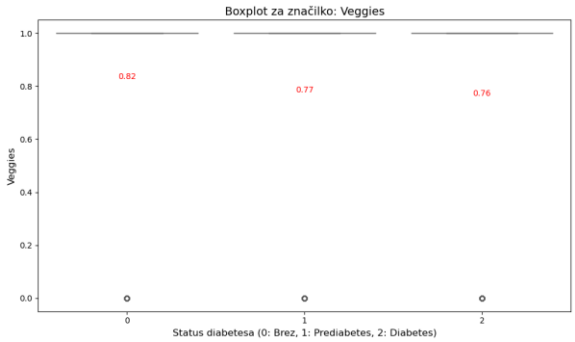
Ali oseba poje 1 ali več sadja na dan.
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da

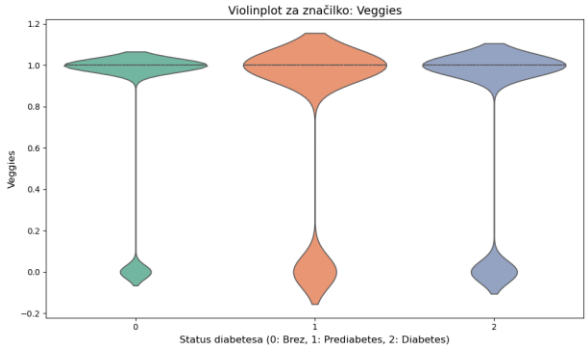




Spremenljivka Veggies

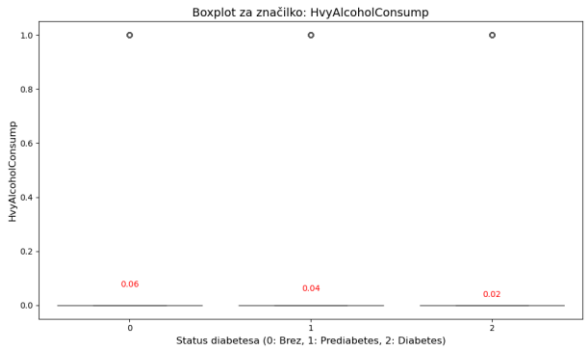
Ali oseba je zelenjavo 1x ali večkrat na dan.
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da

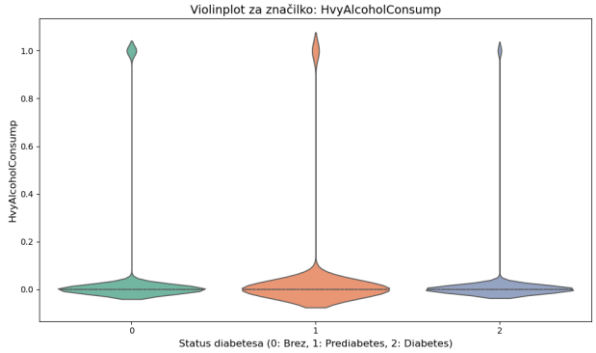




Spremenljivka HvyAlcoholConsump

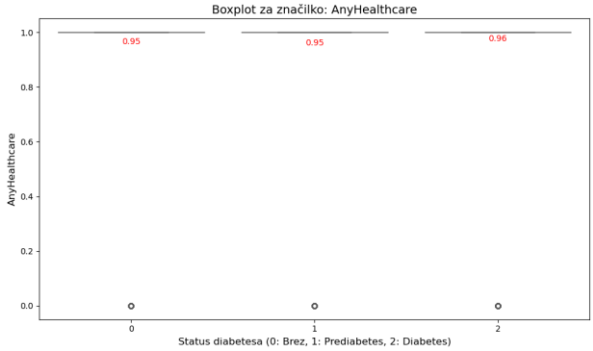
Ali je oseba prekomerni uživalec alkoholnih pijač.
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da

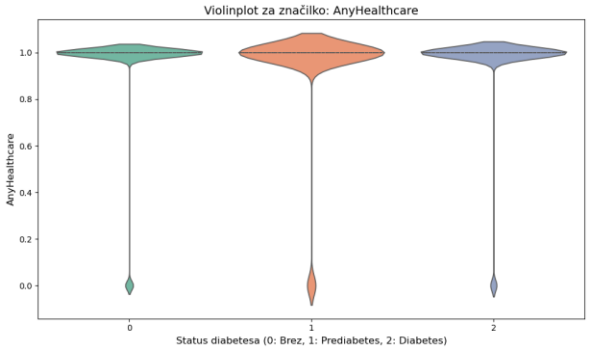




Spremenljivka AnyHealthcare

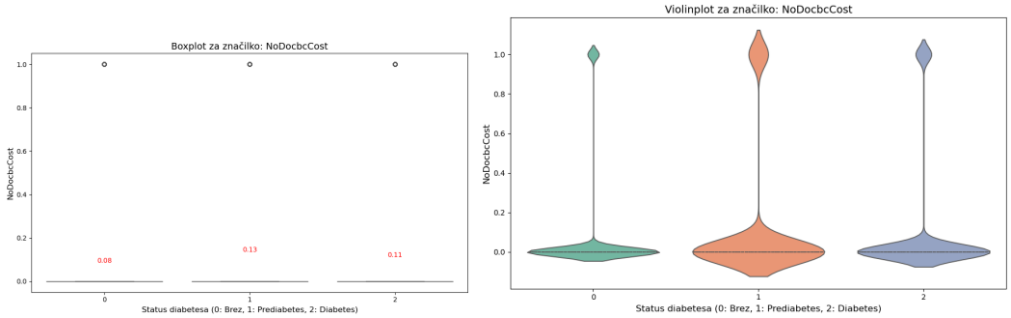
Ali ima oseba katerokoli obliko zdravstvenega zavarovanja.
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da





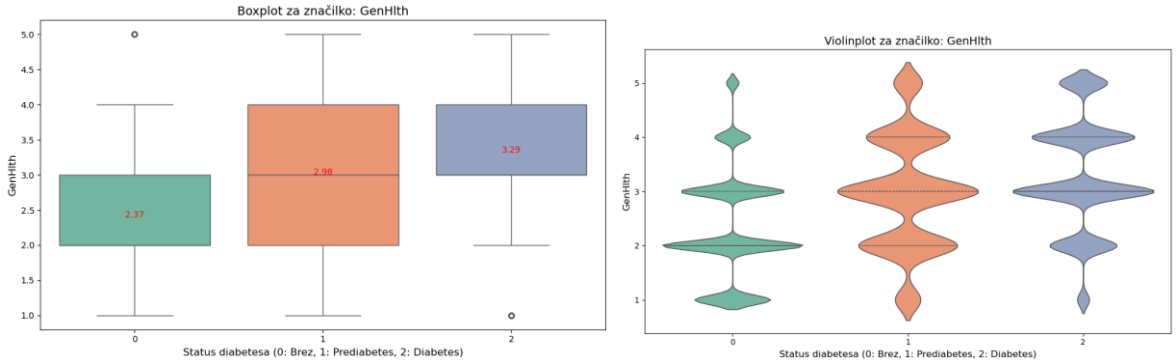
Spremenljivka NoDocbcCost

Ali se oseba ni mogla udeležiti zdravstvenega pregleda v zadnjem letu, ker je cena predraga.
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da



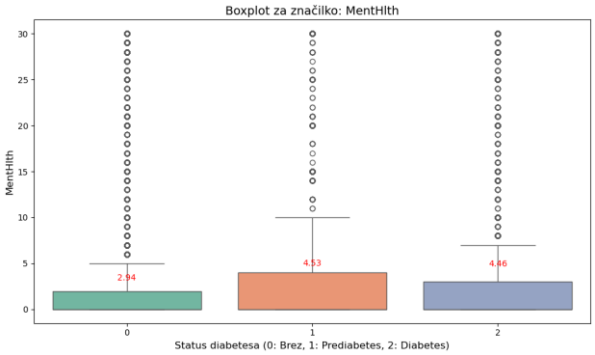
Spremenljivka GenHlth

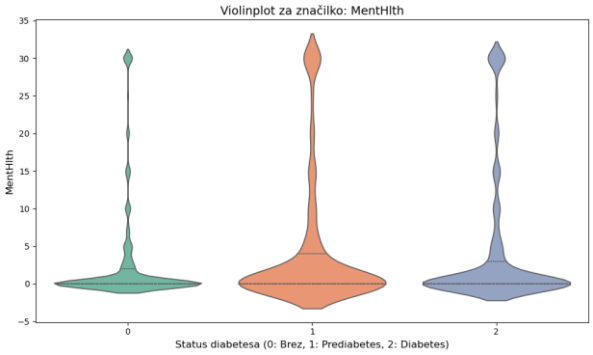
Stopnja s katero je anketirana oseba označila svoje zdravstveno stanje od 1 do 5.
Spremenljivka je ordinalna in številčno kodirana.
1 - Excellent (Odlično)
2 - Very good (Zelo dobro)
3 - Good (Dobro)
4 - Fair (Zadovoljivo)
5 - Poor (Slabo)



Spremenljivka MentHlth

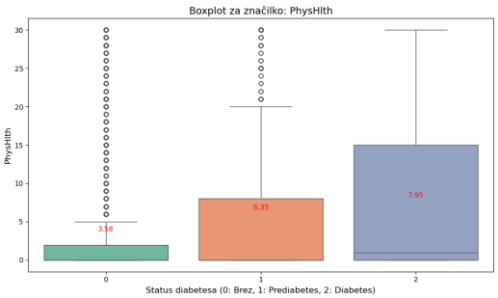
Koliko dni v zadnjih 30 dni od izvedbe ankete je bilo anketirani osebi slabo, kar se tiče mentalnega zdravja.
Spremenljivka je kvantitativna (njene vrednosti so od 0 do 30).

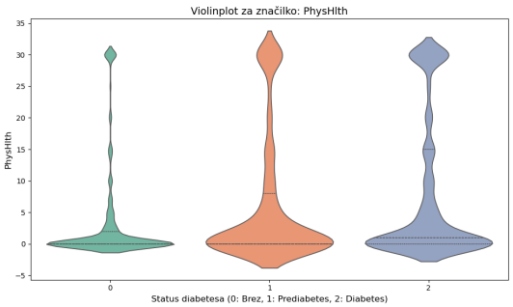




Spremenljivka PhysHlth

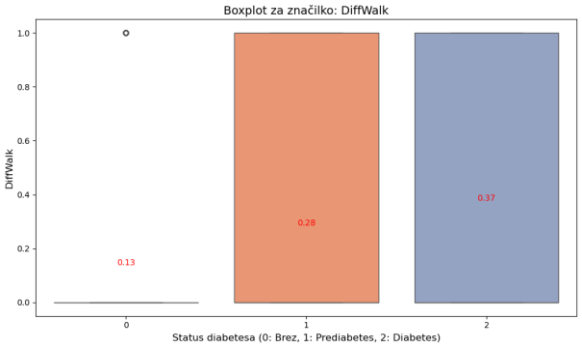
Koliko dni v zadnjih 30 dni od izvedbe ankete je bilo anketirani osebi slabo, kar se tiče fizičnega zdravja (bolezni in poškodbe).
Spremenljivka je kvantitativna (njene vrednosti so od 0 do 30).

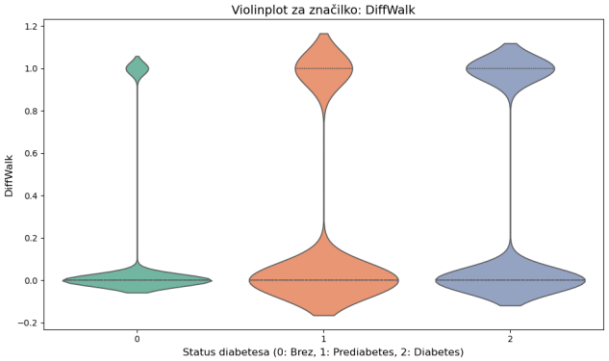




Spremenljivka DiffWalk

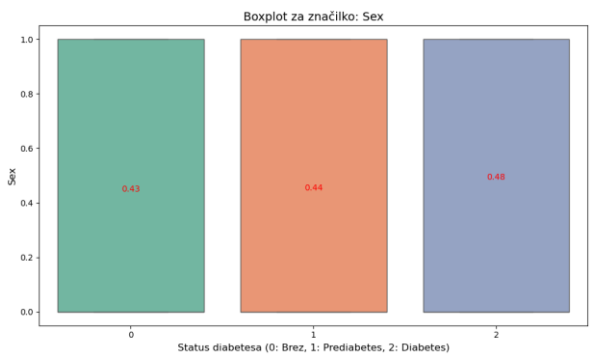
Ali oseba ima probleme s hojo.
Spremenljivka je nominalna in številčno kodirana.
0 – ne
1 – da

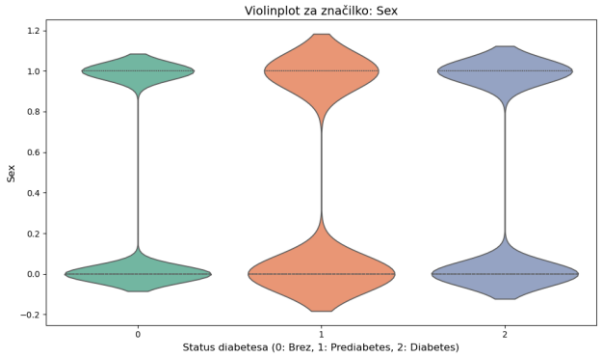




Spremenljivka Sex

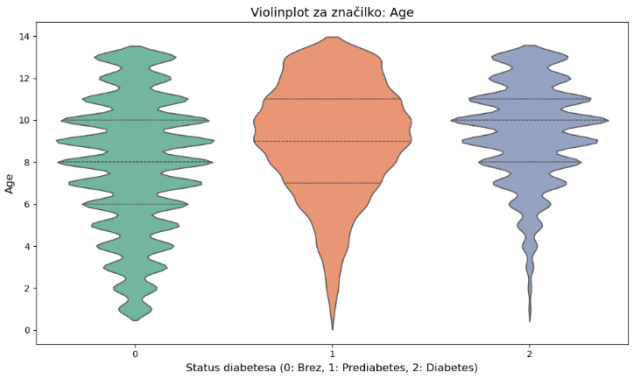
Katerega spola je oseba.
Spremenljivka je nominalna in številčno kodirana.
0 – ženski spol
1 – moški spol

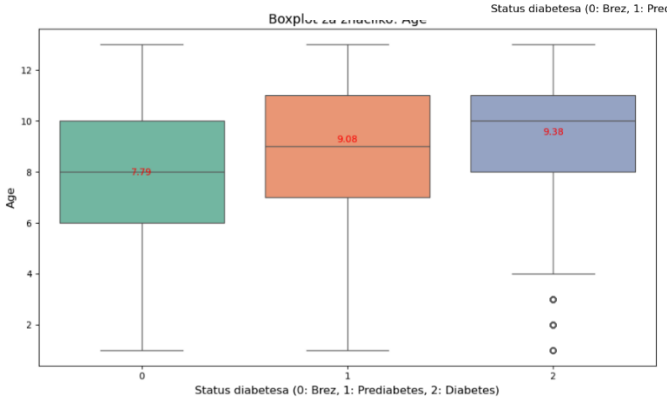




Spremenljivka Age

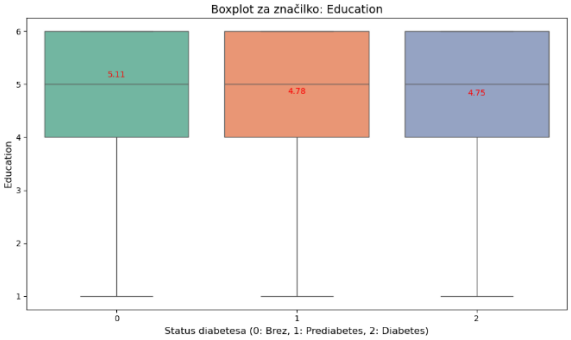
V katero izmed 13 starostnih skupin spada oseba.
Spremenljivka je ordinalna in številčno kodirana.
1 --> 18 <= AGE <= 24
2 --> 25 <= AGE <= 29
3 --> 30 <= AGE <= 34
4 --> 35 <= AGE <= 39
5 --> 40 <= AGE <= 44
6 --> 45 <= AGE <= 49
7 --> 50 <= AGE <= 54
8 --> 55 <= AGE <= 59
9 --> 60 <= AGE <= 64
10 -> 65 <= AGE <= 69
11 -> 70 <= AGE <= 74
12 -> 75 <= AGE <= 79
13 -> 80 <= AGE <= 99





Spremenljivka Education

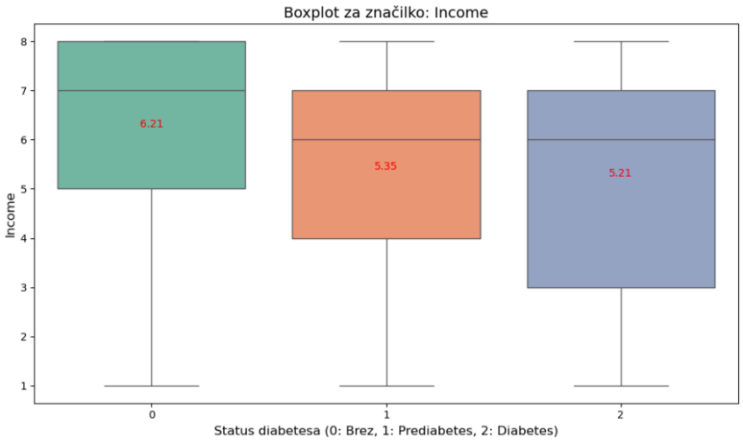
Kateri izmed 6 nivojev odraža doseženo izobrazbo osebe.
Spremenljivka je ordinalna in številčno kodirana.
1 -> Nikoli ni hodil v šolo ali pa bil samo v vrtcu
2 -> Od 1. do 8. razreda (osnovno)
3 -> Od 9. do 11. razreda (nekaj srednje šole)
4 -> 12. razred ali GED (srednješolski maturant)
5 -> Fakulteta od 1 do 3 let (nekaj faksa ali tehnične šole)
6 -> Visoka šola 4 ali več let (višja diploma)



Spremenljivka Income

V kateri izmed 6 dohodkovnih nivojev (celotno gospodinjstvo osebe) spada oseba. Spremenljivka je ordinalna in številčno kodirana.

1 -> Manj kot 10.000$
2 -> 10.000$ do manj kot 15.000$
3 -> 15.000$ do manj kot 20.000$
4 -> 20.000$ do manj kot 25.000$
5 -> 25.000$ do manj kot 35.000$
6 -> 35.000$ do manj kot 50.000$
7 -> 50.000$ do manj kot 75.000$
8 -> 75.000$ ali več



Opis problema 1

Za prvi problem pri tej nalogi smo si izbrali napovedovanje spremenljivke Diabetis_012, oz. ali oseba ima sladkorno bolezen ali ne. Ker je nediagnosticirana sladkorna bolezen tako stroškovna smo se odločili, da bi uporabili podatkovno bazo in s pomočjo njenih podatkov natrenirali dva modela, ki napovedujeta ali oseba ima sladkorno bolezen ali ne. En model, ki smo ga natrenirali je bil logistična regresija, drugi pa RandomForest model. Oba modela smo nato mogli še optimizirati, s ciljem da bi glede na podane podatke napovedovali prisotnost sladkorne bolezeni z boljšo točnostjo. To optimizacijo smo si zadali, da bi naredili tako, da bi iz modela izločili vse spremenljivke, ki nimajo velike pomembnosti pri napovedi sladkorne bolezni glede na dane podatke.

Opis problema 2

Za drugi problem smo izbrali nalogo ustvarjanja priporočil za prediabetike. Cilj je določiti, katere spremenljivke in v kakšni meri jih je treba spremeniti, da bi oseba prešla iz rizičnega stanja prediabetika v zdravo stanje (nediabetik). Ta analiza temelji na uporabi napovednih modelov, kot sta logistična regresija in Random Forest, ki na podlagi obstoječih podatkov simulirata vpliv sprememb posameznih značilk na verjetnost, da oseba postane nediabetik. S tem omogočamo ciljna in informirana priporočila, ki so specifična za posameznika in njihove značilnosti

ANALIZA IN REZULTATI

Ker so bili vsi nominalni in ordinalni podatki že številsko kodirani, ni bilo potrebno veliko urejanja ali čiščenja podatkov. Za prvi problem smo morali izvesti le predprocesiranje celotne podatkovne baze, pri čemer smo odstranili vse osebe, ki so pri spremenljivki Diabetes_012 imele vrednost 1, kar označuje stanje prediabetesa. To smo storili zato, ker nas pri napovedovanju zanima zgolj ločnica med osebo z diabetesom in osebo brez te bolezni.

Z opisanim korakom se je število vzorcev v podatkovni bazi zmanjšalo iz 253.680 na 249.049, kar ni povzročilo bistvene spremembe glede velikosti nabora uporabnih podatkov.

Pri analizi podatkov smo najprej izvedli ustrezne statistične teste. Za nominalne spremenljivke smo uporabili hi-kvadrat test, za ordinalne spremenljivke Kruskal-Wallisov test, za kvantitativne spremenljivke pa t-test in Welchov test. Pred uporabo teh dveh testov smo preverili normalnost porazdelitve s Kolmogorov-Smirnovim testom ter homogenost varianc z Levenovim testom. Ko kvantitativni podatki niso bili normalno porazdeljeni, smo namesto tega uporabili Mann-Whitneyjev test.

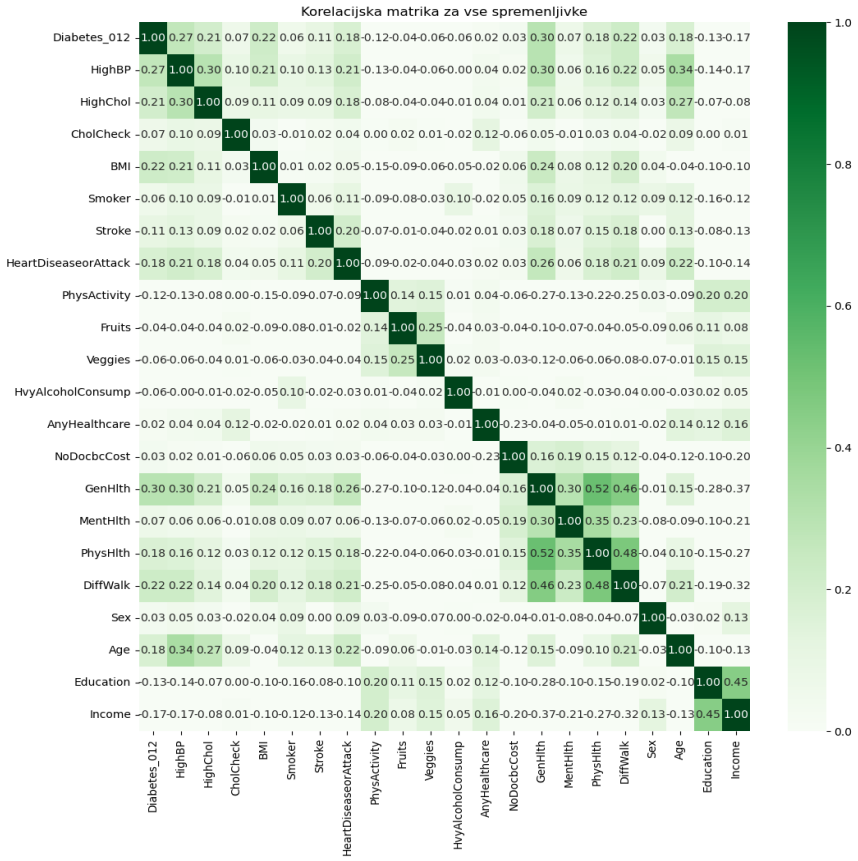
V spodnji tabeli so rezultati testov za vsako spremenljivko v podatkovni bazi, s tem, da je spremenljivka Diabetes_012 uporabljena kot odvisna spremenljivka. Podatki so urejeni glede na p-vrednost naraščujoče.

	Variable	Test	Statistic	P-Value	Significant
0	HighBP	Chi-Square test	1.806262e+04	0.000000e+00	True
18	Age	Kruskal-Wallis test	8.269666e+03	0.000000e+00	True
16	DiffWalk	Chi-Square test	1.249357e+04	0.000000e+00	True
15	PhysHlth	Mann-Whitney U test	2.910964e+09	0.000000e+00	True
13	GenHlth	Kruskal-Wallis test	2.152463e+04	0.000000e+00	True
19	Education	Kruskal-Wallis test	3.787391e+03	0.000000e+00	True
7	PhysActivity	Chi-Square test	3.647181e+03	0.000000e+00	True
20	Income	Kruskal-Wallis test	7.005254e+03	0.000000e+00	True
5	Stroke	Chi-Square test	2.902810e+03	0.000000e+00	True
3	BMI	Mann-Whitney U test	2.331896e+09	0.000000e+00	True
1	HighChol	Chi-Square test	1.053504e+04	0.000000e+00	True
6	HeartDiseaseorAttack	Chi-Square test	8.180575e+03	0.000000e+00	True
2	CholCheck	Chi-Square test	1.085069e+03	5.809239e-238	True
4	Smoker	Chi-Square test	9.635432e+02	1.509376e-211	True
9	Veggies	Chi-Square test	8.405152e+02	8.386730e-185	True
10	HvyAlcoholConsump	Chi-Square test	8.353496e+02	1.113358e-183	True
8	Fruits	Chi-Square test	4.335626e+02	2.725768e-96	True
14	MentHlth	Mann-Whitney U test	3.564682e+09	1.080463e-95	True
12	NoDocbcCost	Chi-Square test	2.733823e+02	2.078555e-61	True
17	Sex	Chi-Square test	2.505281e+02	1.992061e-56	True
11	AnyHealthcare	Chi-Square test	6.547426e+01	5.887806e-16	True

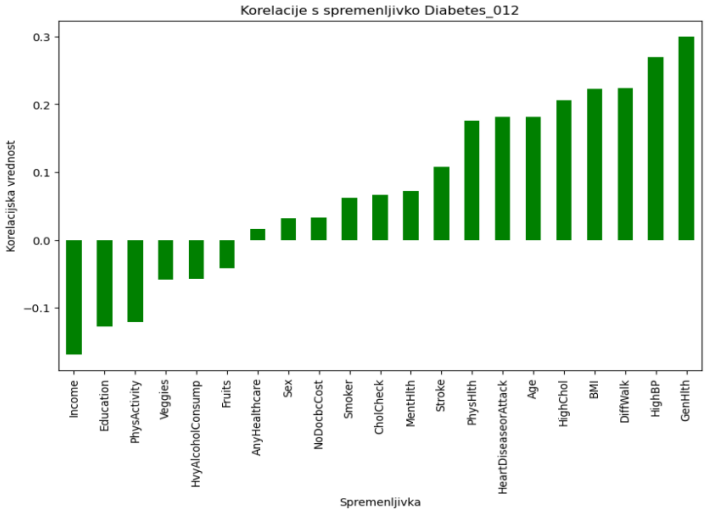
Ker je p-vrednost pri vsaki spremenljivki bila manj kot 0.05 (pravilo P-value < 0.05), pomeni da so vse spremenljivke statistično signifikantne za odvisno spremenljivko Diabetes_012. To pomeni, da je ničelna hipoteza (spremenljivka ni povezana z odvisno spremenljivko) za vsako spremenljivko ovržena in s tem, da ni naključje, da je vsaka posamezna spremenljivka povezana z Diabetes_012 spremenljivko.

Po dobljeni rezultati statistične signifikance, smo dobili hipotezo, da modela za napovedovanje spremenljivke Diabetes_012 ne bo mogoče izboljšati od modela, ki uporablja vse spremenljivke iz podatkovne baze za napoved, če oseba ima sladkorno bolezen.

Še preden pa smo se lotili vzpostavljanja modelov, smo še naredili korelacijsko analizo vseh spremenljivk med sabo, ter izračun VIF (Variance Inflation Factor) faktorjev za vsako od njih. Spodaj je prikazana korelacijska mreža vseh spremenljivk.



Kar se najbolj opazi je, da sta spremenljivki GenHlth in PhysHlth najbolj korellirani med sabo. Z PhysHlth pa še DiffWalk spremenljivka, kar ima vse skupaj dosti smisla, saj oseba pri samooceni zdravja najbolj ocenjuje oz. kritizira svoje fizične sposobnosti, kot je težave s hojo.
Iz tabele smo še prikazali katere spremenljivke pa so najbolj korelirane s spremenljivko Diabetes_012, kar pa prikazuje spodnja slika stolpičnega grafa.



Ker smo pri statističnih testih dobili rezultat, da je vsaka spremenljivka statistično pomembna za napovedovanje diabetesa, smo se še odločili preveriti multikolinearnost med spremenljivkami. S tem smo videli, ali spremenljivke imajo kakšen vpliv ena na drugo. Na podlagi dobljenih rezultatov bi lahko odstranili tiste spremenljivke iz modela, ki imajo veliko povezanost z drugimi spremenljivkami in tako potencialno izboljšali regresijski model. Tega smo se lotili s pomočjo računanja VIF faktorjev.
VIF (Variance Inflation Factor) je mera multikolinearnosti med neodvisnimi spremenljivkami v regresijskem modelu. Multikolinearnost nastane, ko sta dve ali več neodvisnih spremenljivk v modelu močno povezani med seboj, kar lahko povzroči težave pri interpretaciji koeficientov regresijskega modela. Kako si lahko interpretiramo VIF vrednosti:
- VIF ≈ 1: Spremenljivka ni povezana z drugimi, kar je idealno.
- VIF med 1 in 5: Sprejemljivo, nekaj povezave z drugimi spremenljivkami, a ne problematično.
- VIF > 5: Potencialna težava; preveri spremenljivko in njene povezave.
- VIF > 10: Visoka multikolinearnost; spremenljivka zelo verjetno povzroča težave.
Spodnja tabela prikazuje VIF faktorje za vsako spremenljivko:

	Spremenljivka	VIF
19	Education	29.643257
2	CholCheck	22.968166
11	AnyHealthcare	20.875093
3	BMI	18.157011
20	Income	14.252721
13	GenHlth	10.683791
18	Age	9.825327
9	Veggies	5.850786
7	PhysActivity	4.676371
8	Fruits	3.037604
0	HighBP	2.285119
1	HighChol	2.016293
15	PhysHlth	1.994117
4	Smoker	1.929858
17	Sex	1.910840
16	DiffWalk	1.834463
14	MentHlth	1.458271
6	HeartDiseaseorAttack	1.289310
12	NoDocbcCost	1.212497
5	Stroke	1.127038
10	HvyAlcoholConsump	1.083856

Tabela je urejeno padajoče glede na izračunane faktorje. Glede na dobljene rezultate nam je smiselno, da je spremenljivka Education na vrhu, saj je zelo verjetno odvisna od spremenljivk kot sta prihodek (Income) in starost (Age).

Rezultati problema 1

Modela, ki smo si ju izbrali za napovedovanje sladkorne bolezni sta logistična regresija in random forest. Izbrali smo pa ta dva, ker se nam zdita najbolj primerna modela, kot en regresijski in en inteligente, glede na to da gre za klasifikacijski problem. Oba modela smo najprej trenirali na vseh podatkih prirejene podatkovne baze, ter pri obeh uporabili 30% velikost testne množice.
Rezultati logistično regresijskega modela:

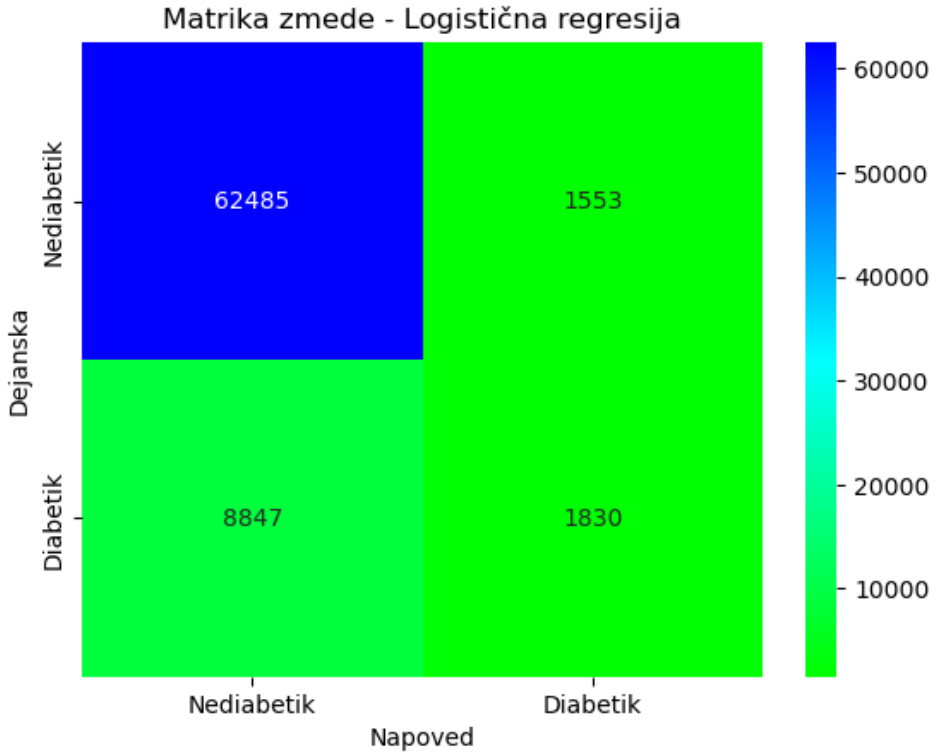
```
Rezultati logistične regresije:
              precision    recall  f1-score   support

           0       0.88      0.98      0.92     64038
           2       0.54      0.17      0.26     10677

    accuracy                           0.86     74715
   macro avg       0.71      0.57      0.59     74715
weighted avg       0.83      0.86      0.83     74715

Točnost modela: 0.8608043900153918
```

Model logistične regresije ima 86% točnost napovedi stanja sladkorne bolezni osebe, po tem ko je bil natreniran z uporabo vseh spremenljivk. Spodaj je še tudi matrika zmede za ta model:



Rezultati random forest modela:

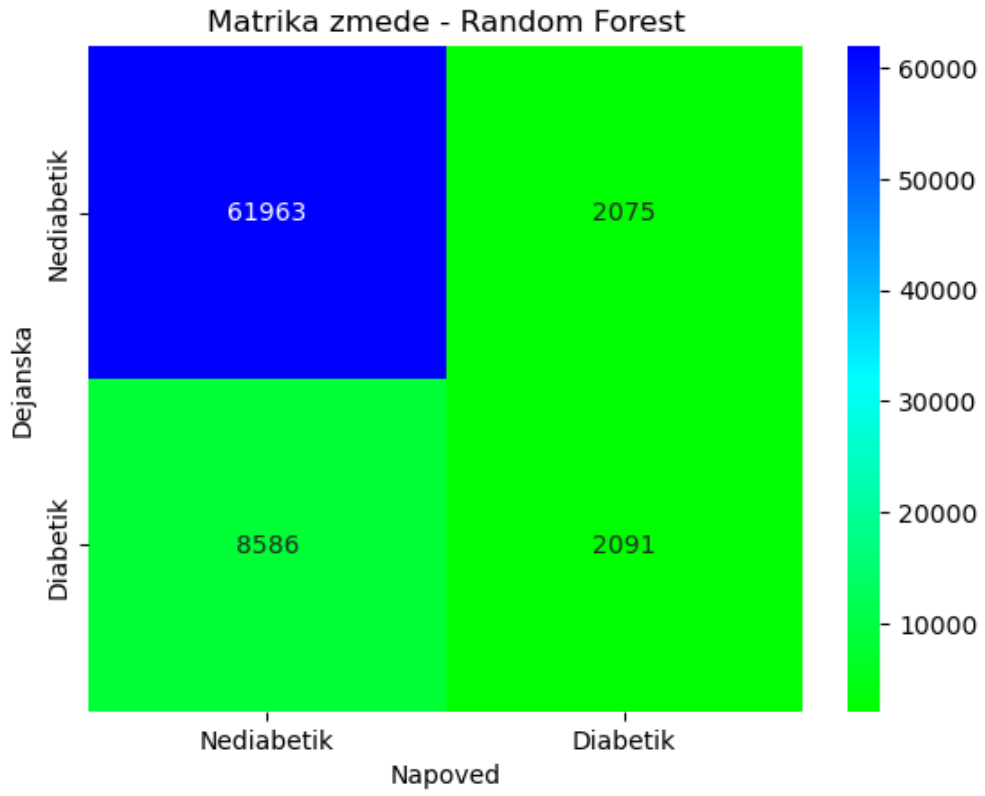
```
Rezultati Random Forest:
              precision    recall  f1-score   support

           0       0.88      0.97      0.92     64038
           2       0.50      0.20      0.28     10677

    accuracy                           0.86     74715
   macro avg       0.69      0.58      0.60     74715
weighted avg       0.82      0.86      0.83     74715

Točnost modela: 0.8573111155725088
```

Random Forest model ima malenkost slabšo točnost napovedi kot regresijski model. Spodaj je prikazana matrika zmede:



Logistična regresija je linearen model za klasifikacijo, ki napoveduje verjetnost, da neka opazovanost pripada določenemu razredu (npr. diabetik ali nediabetik). Model temelji na logistični funkciji (sigmoidni funkciji), ki omogoča, da napovedi ostanejo znotraj intervala [0,1].

Random Forest je ansambelska metoda, ki temelji na številnih odločitvenih drevesih. Namesto enega samega drevesa uporablja več dreves (forest) in združuje njihove napovedi za bolj robusten in natančen rezultat.

Optimizacija problema 1

Optimizacije smo se lotili glede na statistično analizo podatkov.

Najprej smo poskušali natrenirati model glede na rezultate statističnih testov. To pomeni, da smo iz množice podatkov vzeli spremenljivke, ki izstopajo glede na p-value. Te spremenljivke so: CholCheck, Smoker, Veggies, HvyAlcoholConsump, Fruits, MentHlth, NoDocbcCost, Sex in AnyHealthcare. Dobili smo naslednje rezultate:

- Logistic Regression: natančnost modela je padla za 0,02%

```
Rezultati logistične regresije (optimiziran):
              precision    recall  f1-score   support

           0       0.88      0.98      0.92     64038
           2       0.54      0.16      0.25     10677

    accuracy                           0.86     74715
   macro avg       0.71      0.57      0.59     74715
weighted avg       0.83      0.86      0.83     74715

Točnost modela: 0.860683932275982
```

- **Random Forest**: natančnost modela je padla za 0,85%

```
Rezultati Random Forest (optimiziran):
              precision    recall  f1-score   support

           0       0.88      0.95      0.92     64038
           2       0.44      0.23      0.30     10677

    accuracy                           0.85     74715
   macro avg       0.66      0.59      0.61     74715
weighted avg       0.82      0.85      0.83     74715

Točnost modela: 0.8488790738138259
```

Iz rezultatov je razvidno, da nobenega od modelov ne moremo bolj optimizirati, kot pa da uporabimo vse spremenljivke, ki so nam na voljo v podatkovni bazi. Da pa smo se za to prepričali, smo še poskusili optimizacijo na različne načine.
Poskusili smo odstraniti spremenljivke, ki imajo negativno korelacijo z odvisno spremenljivko, ter tiste, katerih vrednost VIF presega 10. Nato smo ponovno odstranili tudi vse spremenljivke z VIF vrednostjo nad 5.
Pri nobenem od teh poskusov nam ni šlo bolj optimizirati nobenega od modelov, kvečjemu smo se le za malenkost oddaljili od točnost modela, ki uporablja vse spremenljivke iz baze. To pa je povsem logično saj glede na prvotni statistični test, so vse spremenljivke statistično signifikantne, kar pomeni da je pomembna za napoved.

Rezultati problema 2

Problema smo se na začetku lotili tako, da smo iz originalne podatkovne baze sfiltrirali vrednosti 2 pri spremenljivki Diabetes_012. S tem je baza imela vrstice samo o nediabetikih in prediabetikih. Nato smo modela logistične regresije in random forest-a natrenirali na teh vrednostih, kjer sta se naučila razpoznavne razlike med nediabetiki in prediabetiki.

Po treniranju modelov smo oba modela uporabili pri simulaciji, ki je služila temu, da smo postopoma spreminjali vsako izmed spremenljivk v podatkovni bazi, dokler model ni oseb, ki so prediabetiki razpoznal kot nediabetik z določeno z verjetnostjo vsaj 80%. V simulaciji smo kot podatkovno bazo v model poslali samo vrstice, ki so osebe zabeležene kot prediabetiki, saj nam je bil cilj najti vrednosti pri posamezniku, ki spremenijo prediabetika v nediabetika.

Primer izpisa pri modelu logistične regresije:

```
[0.2204002 0.7795998]
After improvements [0.84909017 0.15090983] probability
[0.17269064 0.82730936]
After improvements [0.87840598 0.12159402] probability
[0.2250033 0.7749967]
After improvements [0.87412234 0.12587766] probability
[0.32048532 0.67951468]
After improvements [0.8211828 0.1788172] probability
...
[0.62628769 0.37371231]
After improvements [0.94861281 0.05138719] probability
Število priporočil: 4630
Priporočila so bila shranjena v datoteko: prediabetic_recommendations.csv
```

Model nam je najprej izpisal verjetnosti za oba razreda, kjer je prva številka verjetnost, da je oseba nediabetik, druga pa verjetnost, da je oseba prediabetik. Na koncu so se vsa priporočila shranila v prediabetic_recommendations.csv datoteko, kjer piše za čisto vsakega prediabetika, katere spremenljivke mora spremeniti na določeno vrednost, da bo prešel iz stanja prediabetika v nediabetika.

```
After improvements [1. 0.] probability
[0.35 0.65]
After improvements [1. 0.] probability
[0.27 0.73]
After improvements [1. 0.] probability
[0.27 0.73]
After improvements [1. 0.] probability
[0.38 0.62]
After improvements [1. 0.] probability
[0.39 0.61]
After improvements [0.98 0.02] probability
...
[0.37 0.63]
After improvements [1. 0.] probability
Število priporočil: 4631
Priporočila so bila shranjena v datoteko: prediabetic_recommendations_rf.csv
```

Tukaj je še primer izpisa inteligentnega modela random forest, ki dela isto kot model logistične regresije. To je, da prediabetikom išče takšne spremembe v spremenljivkah tako, da bo oseba postala nediabetik (ciljne vrednosti). Isto kot pri logističnem modelu, tudi ta model na koncu naredi csv datoteko v kateri je zapisano za osebo z določenim indeksom v tabeli, katere stvari oz. spremenljivke mora spremeniti, da bo dosegel svoj cilj zdravega življenja.

ZAKLJUČEK

V tej seminarski nalogi smo obravnavali dva povezana problema, pri čemer smo uporabili podatkovno analizo in metode strojnega učenja za napovedovanje ter izboljšanje zdravstvenega stanja posameznikov.

V prvem delu naloge smo se osredotočili na problem klasifikacije, kjer smo z uporabo logistične regresije in Random Forest modelov napovedovali, ali ima posameznik diabetes, prediabetes ali pa je zdrav (nediabetik). Analizirali smo pomembnost različnih značilk, kot so indeks telesne mase (BMI), visok krvni tlak, telesna aktivnost in prehranske navade, ter preučili, kako te vplivajo na tveganje za razvoj diabetesa. Modeli so bili ocenjeni z metričnimi, kot so natančnost, matrike zmede in porazdelitev verjetnosti, kar nam je omogočilo poglobljen vpogled v učinkovitost modelov in njihove napovedne sposobnosti.

V drugem delu naloge smo se osredotočili na prediabetike. Z uporabo istih modelov smo raziskali, katere značilke je treba spremeniti in kako, da bi posameznik prešel iz rizičnega stanja prediabetika v zdravo stanje (nediabetik). Uporabili smo simulacijski pristop, kjer smo iterativno spreminjali vrednosti posameznih značilk in izračunavali, ali te spremembe povečajo verjetnost prehoda v zdravo stanje. Pri tem smo upoštevali omejitve vrednosti značilk (nominalnih, ordinalnih in kvantitativnih), da smo zagotovili

realistično prilagoditev podatkov. Pridobljena priporočila so bila nato shranjena v CSV datoteko, kar omogoča njihovo enostavno interpretacijo in nadaljnjo analizo.

Seminarska naloga je pokazala, kako lahko kombinacija podatkovne analitike in strojnega učenja nudi konkretne in uporabne rešitve za različne zdravstvene probleme. Prvi del naloge je zagotovil natančen sistem za klasifikacijo zdravstvenega stanja, drugi del pa je ponudil personalizirana priporočila za izboljšanje stanja prediabetikov. S tem smo dokazali, da so sodobni podatkovni pristopi ključnega pomena pri reševanju kompleksnih zdravstvenih izzivov in oblikovanju preventivnih ukrepov.