A black and white logo

Description automatically generated

Projekt R

Analiza nutritivnih vrijednosti hrane iz DM-a, McDonalds-a, KFC-a i Burger King-a

Bruno Petković, Ak.god. 2023/2024,

mentor prof. dr. sc. Radovan Zentner

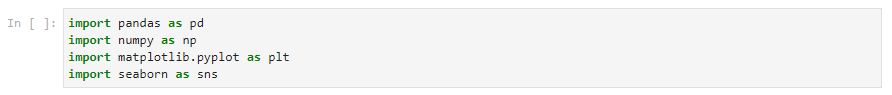
UVOD – Početni problemi i odabir istraživačkih pitanja

Tema projektnog zadatka za Projekt R vezana je za prikupljanje podataka, uređivanje istih i analizu dobivene .csv datoteke. Temeljni problem bio je pronaći potrebne podatke iz vjerodostojnih izvora. Za trgovinu DM zadatak je bio jednostavan zbog pristupačnosti željenih podataka na njihovoj web stranici. Restorani brze hrane - McDonalds, KFC i Bruger King otežali su posao zbog nekonzistentnosti u artiklima i nutritivnom vrijednostima.

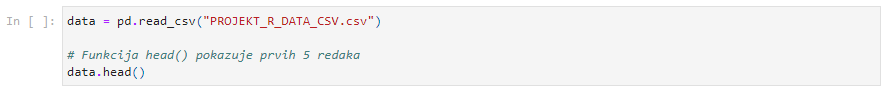
Podatci su izvađeni iz MyFitnessPal aplikacije. MyFitnessPal je svjetski poznata aplikacija koja nudi mogućnost skeniranja bar koda artikla te time dohvaća relevantne podatke nutritivnih vrijednosti vezane za taj artikl. Njihova baza podataka je opširna i pouzdana zbog čega je savršena za ovaj projekt.

Podatci su prikupljani u excel datoteku te su nakon uređivanja i organiziranja exportani u .csv oblik radi daljnje jednostavnosti analize. Analiza i vizualizacija podataka biti će izvršeni u jupyter bilježnici koja koristi jezik python. Završeni rad biti će izvezen u word datoteku.

Za početak je potrebno učitati potrebne biblioteke.



Sada ćemo učitati skup podataka u radno okruženje i dati mu ime "data". Također prije analize je dobra praksa upoznati se sa skupom podataka kroz nekoliko funkcija.



A screenshot of a computer

Description automatically generated

Kada smo se upoznali s načinom formatiranja podataka možemo započeti s istraživačkim pitanjima. Odabrana pitanja:

* Postoji li značajna povezanost između vrste hrane (npr. "zdrava" hrana ili fast-food) i kategorije (npr. "niska količina masti" ili "visoka količina masti")?
* Postoji li značajna povezanost između vrste hrane i kategorije šećera (npr. "niska količina šećera" ili "visoka količina šećera")?
* Je li prosječna energetska vrijednost u DM-u statistički značajno različita od prosječne energetske vrijednosti u fast-food restoranima?
* Postoji li linearna veza između količine masti i energetske vrijednosti u hrani?
* Postoji li statistički značajna razlika u udjelu trans masti između različitih vrsta fast-food restorana (npr. McDonald's, KFC, Burger King)?
* Može li se na temelju nutritivnih vrijednosti proizvoda iz različitih restorana modelirati sustav koji pogađa iz kojeg restorana proizvod dolazi?

PITANJE 1: Postoji li značajna povezanost između vrste hrane (npr. "zdrava" hrana ili fast-food) i kategorije (npr. "niska količina masti" ili "visoka količina masti")?

Odabiremo Chi-kvadrat test kako bi dobili količinu različitosti varijanci količine masti u zadanim kategorijama. Chi-kvadrat test je statistički test koji se koristi kada treba odrediti povezanost više kategorijskih varijabli. U našem slučaju to su kategorije „vrsta trgovine“ i „količina masti“.

Postavljamo hipoteze:

Nulta hipoteza (H0): Vrsta trgovine i količina masti su nezavisne.

Alternativna hipoteza (H1): Vrsta trgovine i količina masti su zavisne

Provodimo Chi-kvadrat test u jupyter bilježnici.

A screenshot of a computer program

Description automatically generated

Dobiveni rezultati pokazuju kako ne možemo odbaciti nultu hipotezu u korist alternativne jer P-vrijednost nije manja od 0.05 (po dogovoru).

No pokrenut ćemo drugi test - Fisherov egzaktni test. Ne možemo se osloniti na Chi-kvadrat test kod manjih skupova podataka, zato koristimo Fisherov egzaktni test.

A screenshot of a computer code

Description automatically generated

Kako je P-vrijednost ovaj put manja od 0.05 možemo odbaciti H0 u korist H1.

Rezultati pokazuju kako postoji značajna povezanost između vrste hrane (Fast Food / DM) i kategorije masti (visoka / niska). Omjer Šansi = 16 pokazuje da je šansa za visoku razinu masti u Fast Food hrani 16 puta veća nego kod hrane iz DM-a.

A screen shot of a computer code

Description automatically generated

A screenshot of a graph

Description automatically generated

PITANJE 2: Postoji li značajna povezanost između vrste hrane i kategorije šećera (npr. "niska količina šećera" ili "visoka količina šećera")?

Ovo istraživačko pitanje je po procesu slično prošlome pa nije potrebno objašnjavati proces dobivanja podataka.

A screen shot of a computer code

Description automatically generated

Pretragom interneta saznajemo da je preporučena doza šećera po 100g proizvoda manja od 10g, ali niti jedan artikl ne prelazi tu granicu. Stoga ćemo staviti granicu visokog i niskog šećera na lokalnim podatcima na 3g. Rezultati takvog testa pokazuju da nema dovoljno statističkih dokaza za odbacivanje nulte hipoteze tj. ne postoji značajna korelacija između kategorije hrane i razine šećera.

A computer screen shot of a computer code

Description automatically generated

A graph with a green and pink rectangles

Description automatically generated with medium confidence

Vidimo kako kod Fast Food restorana ima više šećera, ali na ovolikom uzorku nema statistički značajno više kako bi mogli nešto zaključiti.

PITANJE 3: Je li prosječna energetska vrijednost u DM-u statistički značajno različita od prosječne energetske vrijednosti u fast-food restoranima?

Kao početak odgovora istaknuo bi jedan od prijašnjih odgovora - razlika u masti. Mast je makronutrijent koji uvelike pridonosi energetskoj vrijednosti hrane, te na temelju tog pitanja možemo ponuditi odgovor i na ovo, ali tada bi zanemarili ostale faktore koji ulaze u cjelovitu energetsku vrijednost.

Nulta hipoteza glasi: Ne postoji razlika u prosjecima energetskih vrijednosti hrane iz Fast Food restorana i DM-a.

Provest ćemo T-test za nezavisne uzorke.

A screen shot of a computer code

Description automatically generated

Kako je P-vrijednost jako mala (znatno ispod 0.05), možemo zaključiti kako postoji značajna razlika u prosječnoj energetskoj vrijednosti između dvije kategorije hrane. Također T-statistika = -6.89 pokazuje kako je energetska vrijednost u DM-u niža od Fast Food restorana.

Vizualizirajmo dobivene rezultate kako bi bolje prikazali jaz među energetskom vrijednosti. Napravit ćemo Box-plot koji prikazuje medijan, grupaciju podataka i outliere.

A screenshot of a computer program

Description automatically generated

A graph with a bar chart

Description automatically generated with medium confidence

Uočavamo kako je artikl s najvišom energetskom vrijednosti kod DM-a u istoj energetskoj razini kao i najnižih 25% artikala Fast Food restorana.

PITANJE 4: Postoji li linearna veza između količine masti i energetske vrijednosti u hrani?

Prethodni odgovor navodi nas na korelaciju između količine masti i energetske vrijednosti u hrani. Testirajmo tu hipotezu.

A computer screen shot of a computer code

Description automatically generated

A graph with a red line and blue dots

Description automatically generated

Pokrenuli smo funkciju linearne regresije koja pokazuje trend kretanja energetske vrijednosti u ovisnosti o količini masti. Na tim parametrima smo također pokrenuli Pearsonov test koeficijenta korelacije kako bi našli koliko su pojmovi međusobno povezani.

Pearsonov koeficijent korelacije je 0 ako nema povezanosti, -1 kad je negativna povezanost, a 1 kad je pozitivna povezanost (rast jedne varijable je direktno povezan s rastom druge). Dobiveni koeficijent je 0.87 što je blizu 1, dakle statistički možemo zaključiti kako su količina masti i energetska vrijednost međusobno povezane vrijednosti. Znatiželje radi provest ćemo Pearsonov test na drugim varijablama iz tablice kako bi dočarali koje to varijable ne utječu na energetsku vrijednost.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

Podsjećam, P-vrijednost veća of 0.05 pokazuje da količina podataka na kojima je test zasnovan nije dovoljno velika kako bi se statistički došlo do zaključka, stoga kategoriju "Vlakna" i "Sol" ne možemo uzeti u obzir.

PITANJE 5: Postoji li statistički značajna razlika u udjelu trans masti između različitih vrsta fast-food restorana (npr. McDonald's, KFC, Burger King)?

Trans masti su česta tema zdravstvenih blogova zadnjih godina. Dokazano je kako je njihova konzumacija povezana sa začepljivanjem krvnih žila kroz proces povećavanja LDL kolesterola ("loš kolesterol", začepljuje žile) i istovremenog smanjivanja količine HDL kolesterola ("dobar kolesterol", čisti žile).

Nas zanima postoji li razlika u količini trans masti po 100g proizvoda iz odabranih Fast Food restorana. DM nećemo uzeti u obzir jer je očito da iz zadanih podataka niti jedan proizvod nema prisutnih trans masti.

Provest ćemo ANOVA test. ANOVA test je statistička tehnika koja se koristi za testiranje razlika srednjih vrijednosti unutar grupe od tri ili više člana. Uspoređuju se varijance unutar grupe te se utvrđuje jesu li razlike veće od onoga što bismo očekivali zbog slučajnosti.

Nulta hipoteza (H0): Srednje vrijednosti trans masti su jednake za sve vrste Fast Food restorana.

A screenshot of a computer code

Description automatically generated

Kako je P-vrijednost = 0.023 odbacujemo H0 i zaključujemo da se jedan član grupe ističe.

Da bi otkrili koji je to provest ćemo Tukey-ev test.

A screenshot of a computer code

Description automatically generated

Tukey-ev test uparuje sve moguće parove te ih zasebno testira. Dobiveni rezultati pokazuju kako se Burger King i KFC najviše razlikuju. Interpretirajući rezultate zaključujemo kako najmanju količinu trans masti po prosjeku ima KFC, McDonalds je u sredini i Burger King ima najviše.

Prikažimo podatke vizualno.

A screenshot of a computer program

Description automatically generated

A graph with different colored squares

Description automatically generated

PITANJE 6: Može li se na temelju nutritivnih vrijednosti proizvoda iz različitih restorana modelirati sustav koji pogađa iz kojeg restorana proizvod dolazi?

Najzanimljivije pitanje ovog projekta dolazi na kraju - izrada modela za predviđanje. Za početak ćemo obraditi skup podataka tako da je ostalo samo najpotrebnije za logističku regresiju. Ukloniti ćemo stupce poput "IME PROIZVODA", "IME TRGOVINE"... odvojiti ćemo podatkovni skup na način da su od svakog restorana po 9 artikla na kojim se trenira model i zadnji artikl na kojem će se testirati.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

Naš model je točno pogodio ime restorana u 87.5% slučajeva na temelju samo 10 artikla iz 4 restorana.

Sporni dio je između Burger King-a i McDonalds-a jer kako smo vidjeli u prethodnim testovima to su dva relativno bliska restorana.

ZAKLJUČAK

Ovaj istraživački projekt dao nam je uvid u nutritivne vrijednosti proizvoda iz trgovine DM te restorana brze hrane poput McDonald's-a, KFC-a i Burger King-a. Kroz analizu podataka i provedbu statističkih testova, došli smo do nekoliko važnih zaključaka:

* Povezanost između vrste hrane i kategorije masti postoji.
* Povezanost između vrste hrane i kategorije šećera postoji ali je količina šećera u granicama.
* Razlika u prosječnoj energetskoj vrijednosti između DM-a i fast-food restorana postoji, hrana iz DM-a je znatno manje klaorična.
* Veza između količine masti i energetske vrijednosti je linearna s jakim koeficijentom korelacije od 0.87.
* Razlika u udjelu trans masti među fast-food restoranima postoji, Burger King hrana sadrži najviše trans masti.
* Razvili smo model koji pogađa iz kojeg je restorana hrana na temelju danih nutritivnih vrijednosti s preciznošću od 87.5%.