**TEHNIČKO VELEUČILIŠTE U ZAGREBU**

**POLITEHNIČKI SPECIJALISTIČKI DIPLOMSKI**

**STRUČNI STUDIJ**

**SPECIJALIZACIJA INFORMATIKA**

**NOSQL I NAPREDNA BIG DATA ANALITIKA**

Tin Milaković

KOMENTIRANJE PROJEKTA 3 – BAYESOVE MREŽE

Zagreb, veljača, 2024.

Sadržaj

[1. Uvod 3](#_Toc158914950)

[2. Najbitnije karakteristike modela Bayesovih mreža 4](#_Toc158914951)

[3. Objašnjavanje modela 5](#_Toc158914952)

[3.1. Prvi pokušaj generiranja modela 5](#_Toc158914953)

[3.2. Drugi pokušaj generiranja modela 5](#_Toc158914954)

[3.3. Treći pokušaj generiranja modela 6](#_Toc158914955)

[4. Prikaz dodatnih podataka dobivenih pri validaciji 8](#_Toc158914956)

[5. Scenariji osjetljivosti 11](#_Toc158914957)

[6. Zaključak 13](#_Toc158914958)

Sadržaj slika

[Slika 1 Model Bayesovih mreža 4](#_Toc158914940)

[Slika 2 Pozadinsko znanje prije kreiranja modela 6](#_Toc158914941)

[Slika 3 Korelacije generiranih veza 7](#_Toc158914942)

[Slika 4 Točnost podataka 8](#_Toc158914943)

[Slika 5 Konfuzijska matrica 8](#_Toc158914944)

[Slika 6 ROC krivulja za ciljnu varijablu vrijednosti „DA“ 9](#_Toc158914945)

[Slika 7 ROC krivulja za ciljnu varijablu vrijednosti „NE“ 10](#_Toc158914946)

[Slika 8 Scenarij sa rasponima najvećih WoE vrijednosti 11](#_Toc158914947)

[Slika 9 Scenarij sa rasponima najmanjim WoE vrijednosti 12](#_Toc158914948)

[Slika 10 Scenarij sa nasumično odabranim rasponima 12](#_Toc158914949)

# Uvod

Prije svega, volio bih komentirati svoj dataset. Poanta danog dataseta je da se vidi kako će pojedina svojstva utjecati na kritičnu temperaturu superkonduktora, no prikupljena svojstva nisu jedinstvena poput „širina“ i „dužina“ već su nekakve iteracije samih svojstava poput „srednja\_vrijednost\_širine“, „standardna\_devijacija\_širine“ i slično tako da se ne može direktno doći do nekakvog korisnog zaključka. Ciljna varijabla je kritična temperatura koju konstantno promatramo što se sa njom dešava modificiranjem svakog svojstva. Vrijednosti same ciljne varijable je iznos temperature u kelvinima (K), no za kategorizaciju je bilo uzeto da, ako je temperatura manja od 30K, spada u kategoriju „DA“ jer takvi superkonduktori spadaju u tip 1 superkonduktora.

Postojalo je puno problema pri kreiranju same mreže. Za uspješan model je uzeto 2000 redaka i 14 stupaca od kojih su 7 nasumičnih i 7 sa najboljim IV (engl. *Information Value*) te ciljna varijabla. Pri diskretizaciji podataka bile su korištene odokativne WoE (engl. *Weight of Evidence*) grupacije od svih podatka, ne samo od 2000 podataka jer nije htjelo raditi sa diskretizacijom po WoE grupacijama od 2000 podataka. Sve je ovo bilo urađeno jer se program konstantno gasio bez ikakvih *loggova* o tome što je uzrokovalo tome. Detaljniji pokušaji kreiranja mreže su objašnjeni u poglavlju o objašnjavanju dobivene mreže.

# Najbitnije karakteristike modela Bayesovih mreža

Grafička reprezentacija – koriste usmjerene cikličke grafove za predstavljanje uvjetnih ovisnosti među varijablama tako što čvorovi predstavljaju varijable, a strelice predstavljaju uvjetne ovisnosti.

Učenje iz podataka – modeli mogu učiti iz podataka korištenjem različitih algoritama, kao što su Bayesian Search, PC, Greedy Thick Thinning, … (primjeri izvučeni iz GeNIe alata)

Predviđanje – mogu se koristiti za predviđanje vjerojatnosti različitih konfiguracija varijabli na temelju novih podataka

Skalabilnost – modeli se mogu koristiti za modeliranje velikih problema s velikim brojem varijabli

Integracija s drugim modelima – Bayesovi modeli se mogu lako integrirati s drugim modelima, poput neuronskih mreža

Izgled modela Bayesovih mreža koristeći alat GeNIe:

A computer screen shot of a computer network

Description automatically generated with medium confidence

Slika 1 Model Bayesovih mreža

# Objašnjavanje modela

Set podataka koji je korišten ima 21263 redaka podataka sa 82 stupca te se za kreiranje mreže nije mogao cijeli iskoristiti. Postoji nekoliko varijacija pokušaja, no samo je jedan donio uspješan rezultat. Pokušaji su odvojeni i dodatno objašnjeni.

## Prvi pokušaj generiranja modela

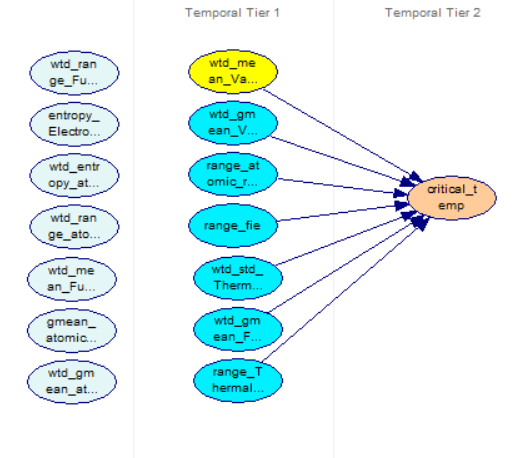
Prvi pokušaj je bio koristiti sve podatke i sve stupce koristeći one sa vrlo visokim IV da utječu na ciljnu varijablu (35 varijabli je zadovoljavalo uvjet gdje je IV > 1.8). Ovaj pokušaj je konstantno dovodio do pada generiranja modela gdje samo generiranje nije prošlo više od 0% i nije potrajalo duže od 30 sekundi. Sa ovim informacijama smo došli do zaključka da je potrebno reducirati ulazni broj podataka te stupaca uzetih za generiranje modela. Za ovaj pokušaj postoje varijacije gdje su uzimane različiti broj grupa za diskretizaciju (5% i 10%) te su probane tehnike poput preimenovanja varijabli u nešto kraće, testirajući hoće li to utjecati na generiranje modela.

## Drugi pokušaj generiranja modela

Drugi pokušaj je koristio samo 14 varijabli od kojih su samo 7 utjecali na ciljnu varijablu. Navedene varijable su imale najveće IV te su iz istog razloga bile odabrane da utječu na ciljnu varijablu dok je ostalih 7 varijabli uzeto nasumično. Odabrano je nasumičnih 4000 redaka kako bi se probao generirati model. U ovom pokušaju bile su, također, varijacije pri generiranju modela gdje se ponovno probala tehnika sa preimenovanjem i različitim brojem grupa za diskretizaciju. Ovaj pokušaj je, nažalost, bio neuspješan, no generiranje modela je bilo uspješnije sa nekoliko pokušaja koji su došli do 50% gdje se sam program konstantno srušio pri svakom novom pokušaju što je dovelo do zaključka da je možda problem u ponovnom prevelikom broju podataka.

## Treći pokušaj generiranja modela

Treći pokušaj je bio postavljen isto kao i drugi, no ovaj put je korišteno samo 2000 redaka od mogućih 21263, odnosno 9.4% podataka. Korištene su WoE grupacije od svih podataka jer nije htjelo proraditi sa grupacijama od 2000 redaka. Generiranjem mreže na ovaj način je uspješno dobiven model po prvi put te se cijeli projekt nakon uradio koristeći ovaj model. Za generiranje modela bilo je potrebno 306 sekundi. Za učenje modela je bilo korišteno pozadinsko znanje koje je prikazano idućom slikom:



Slika 2 Pozadinsko znanje prije kreiranja modela

Zanimljivo je kako je dobiven velik broj novih veza između varijabli koje smo postavili da ne utječu na ciljnu i onih koje utječu. Generiranjem modela se dobilo 15ak novih veza, no ne postoji jedna nova koja utječe na ciljnu varijablu već su većinom generirane između navedenih varijabli sa iznimkama gdje su se veze stvorile između varijabli koje utječu na ciljnu. Koristeći dodatni kod, moguće je provjeriti korelacije između tih varijabli te su dobivene korelacije između ciljnih varijable prikazane idućom slikom:

A black and white text on a white background

Description automatically generated

Slika 3 Korelacije generiranih veza

Kao što se može primijetiti, korelacije između varijabli gdje su kreirane nove veze spadaju u jake korelacije što je Bayesova mreža pri treniranju prepoznala te, vrlo impresivno, dodala nove veze.

# Prikaz dodatnih podataka dobivenih pri validaciji

Prilikom validacije mreže dobijemo nekoliko dodatnih podataka o našoj mreži koji su prikazani idućim slikama:

A black and white text with numbers

Description automatically generated

Slika 4 Točnost podataka

Točnost predstavlja koliko je mreža točna pri predikciji rezultata. Na slici možemo vidjeti kako je mreža bila točna sveukupno 93.05% odnosno u 1861 od 2000 slučajeva te možemo vidjeti i pojedinačne slučajeve zasebno.

A screenshot of a computer

Description automatically generated

Slika 5 Konfuzijska matrica

Konfuzijska matrica nam prikazuje stvarna u odnosu na predviđena stanja. Dijagonalne ćelije nam prikazuju točno klasificirane slučajeve (podebljanje ćelije) dok nam ne dijagonalne ćelije prikazuju netočno klasificirane slučajeve. Matrica nam vizualnije prikazuje točnost predviđanja mreže.

Idućim slikama su prikazane ROC (engl. *Receiver Operating Characteristic*) krivulje:

A screen shot of a graph

Description automatically generated

Slika 6 ROC krivulja za ciljnu varijablu vrijednosti „DA“

ROC krivulja prikazuje *True Positive Rate* (TPR) (osjetljivost) u odnosu na *False Positive Rate* (FPR) (1 - specifičnost) na različitim pragovima predviđanja. Idealna krivulja „grli“ lijevi gornji korner, postižući visoku osjetljivost s malim brojem lažno pozitivnih rezultata što je na navedenoj slici za ciljnu varijablu vrijednosti „DA“ vrlo približno, priča je obrnuta ako je krivulja bliža dijagonali. Koristeći ROC krivulju možemo doći do zaključka kakva je sposobnost mreže da klasificira podatke. Idućom slikom je prikazano stanje „NE“ ciljne varijable te njegova ROC krivulja.

A screen shot of a graph

Description automatically generated

Slika 7 ROC krivulja za ciljnu varijablu vrijednosti „NE“

# Scenariji osjetljivosti

Scenariji osjetljivosti su načini da testiramo kako će promjena vrijednosti varijabli utjecati na drugu varijablu, u našem slučaju će se gledati samo ciljna varijabla i kako utjecaj drugih varijabli se odražava na nju. U scenariju osjetljivosti uzete su svi rasponi gdje pripadaju najveće WoE vrijednost u usporedbu na ciljnu varijablu. Slika tog scenarija je iduća:

A computer screen shot of a computer

Description automatically generated with medium confidence

Slika 8 Scenarij sa rasponima najvećih WoE vrijednosti

Koristeći sve raspone gdje im je WoE vrijednost najveća dobivamo rezultate da je 75% šansa da rezultat spada pod kategoriju „DA“ dok ostatak spada pod kategoriju „NE“. Koristeći ovakve scenarije mi se zapravo „igramo“ sa našom mrežom koristeći bilo kakve kombinacije da vidimo kako će utjecati na izlaz. Za drugi scenarij će se koristiti obrnuta metoda od prve gdje će se uzeti svi rasponi sa najmanjim WoE vrijednostima. Slika tog scenarija je iduća:

A computer screen shot of a computer

Description automatically generated

Slika 9 Scenarij sa rasponima najmanjim WoE vrijednosti

Koristeći sve raspone gdje im je WoE vrijednost najmanja dobivamo rezultate da je 97% da ta vrijednost spada u kategoriju „NE“. Za zadnji scenarij osjetljivosti uzete su nasumične vrijednosti što je rezultiralo da je ciljna varijabla podijeljena točno po sredini, 50% „DA“ i 50% „NE“. Scenarij je prikazan na idućoj slici:

A computer screen with many different colored squares

Description automatically generated with medium confidence

Slika 10 Scenarij sa nasumično odabranim rasponima

# Zaključak

Ovim projektom sam bio više upućen u svijet data analitike i kako nam Bayesova mreža može pomoći sa našim podatcima. Za ovaj projekt je bio preduvjet da se izračunaju WoE i IV od uzetih svojstava koji su napravljeni u prethodnom projektu.

Koristeći dobivene podatke mi dodatno dobivamo informacije koje su varijable korisnije prilikom klasifikacije tako da možemo predati nekakvo pozadinsko znanje prije treniranja mreže. Naravno dobiveni podatci nam pomažu i pri pripremi podataka i samoj interpretaciji podataka.

Koristeći Bayesove mreže možemo vizualizirati kako varijable utječu jedna na drugu te kako se ciljna varijabla mijenja na najmanju promjenu mreže.

Za ovaj projekt se koristio program GeNIe koji je definitivno otežavao, ali i olakšao rješavanje projekta. Ovo je također 3. i zadnji projekt predmeta „No SQL i big data analitika“ te smatram da su projekti vrlo solidan uvod u svijet big data analitike.