Univerzitet u Nišu

ELEKTRONSKI FAKULTET

Katedra za računarstvo

Šum u podacima

SEMINARSKI RAD

Studijski program: Veštačka inteligencija i mašinsko učenje

Student:

Milica Jovanović 1636

Niš, Februar 2024.

Sadržaj

Uvod

Podaci predstavljaju temelj za donošenje odluka u različitim oblastima, od poslovanja i naučnih istraživanja do tehnologije i društvenih nauka. Jedan od ključnih problema u radu sa prikupljenim podacima jeste prisustvo šuma. Šum se može definisati kao nepoželjna ili irelevantna informacija koja može narušiti proces analize, otežati prepoznavanje obrazaca i dovesti do pogrešnih zaključaka. Upravo zbog toga, razumevanje prirode šuma i njegovih efekata postaje sve važniji aspekt u oblasti obrade podataka. Kako je obrada velikih količina podataka postala neophodna za napredak u mnogim oblastima, upravljanje šumom u podacima dobija sve veću važnost.

Šum može dovesti do ozbiljnih problema u analizi, kao što su prekomerna složenost modela, pogrešna klasifikacija, pa čak i potpuno netačne prognoze. Šum u podacima može nastati iz različitih izvora, kao što su greške pri unosu podataka, tehničke smetnje prilikom prikupljanja informacija, ili čak prirodne varijacije koje nisu povezane s promenljivom koja se ispituje. Bez obzira na poreklo, prisustvo šuma može značajno smanjiti tačnost analitičkih modela, uzrokujući da rezultati ne odražavaju stvarnu situaciju. Ovo je posebno problematično u situacijama kada se koriste složeni modeli veštačke inteligencije i mašinskog učenja, gde čak i mala količina šuma može imati velike posledice na performanse modela.

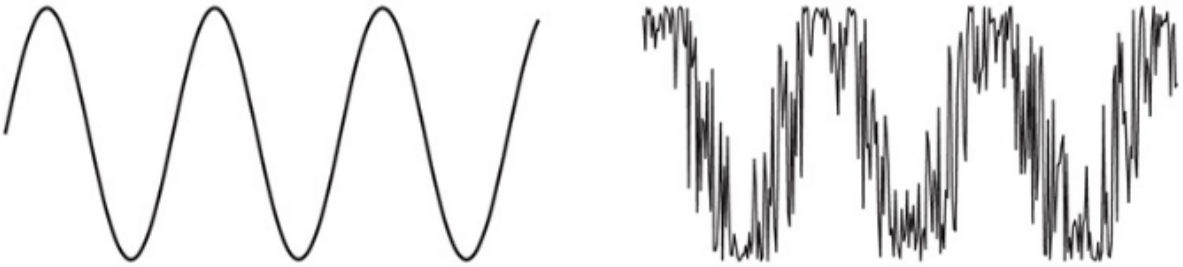
Takođe, šum u podacima može imati značajan uticaj na validnost i pouzdanost istraživačkih rezultata. U društvenim naukama, na primer, prisustvo šuma može dovesti do netačnih zaključaka o ponašanju ili stavovima ispitanika, što može rezultirati neprikladnim preporukama. Slično tome, u prirodnim naukama, šum može ometati identifikaciju stvarnih uzročnih veza među promenljivama, što može dovesti do lažnih pozitivnih ili negativnih rezultata. Stoga, istraživači moraju biti pažljivi u pogledu kvaliteta podataka koje koriste, kao i metoda koje primenjuju za prepoznavanje i uklanjanje šuma.

U ovom seminarskom radu, biće istraženi različiti oblici šuma, njegovi uzroci i načini na koje može uticati na analizu podataka. Pored toga, biće predstavljeni pristupi i tehnike koje se koriste za smanjenje šuma i poboljšanje kvaliteta podataka, kako bi se postigli što tačniji i pouzdaniji rezultati. Kroz analizu cilj je da se prikaže važnost pravilnog tretmana šuma u podacima i da se ukaže na metode koje mogu pomoći u minimiziranju njegovog negativnog uticaja na istraživanja i donošenje odluka.

1. Šta je šum

Podaci iz stvarnog sveta nikada nisu savršeni i često sadrže određenu količinu anomalija ili šuma koji mogu da naškode interpretaciji podataka, izgradnji modela i donošenju odluka modela u mašinskom učenju.

Šum je nasumična komponenta greške merenja. Obično uključuje distorziju vrednosti ili dodavanje lažnih podataka. Na slici 1 prikazana je vremenska serija pre i nakon što je poremećena slučajnim šumom. Ako bi se vremenskoj seriji dodalo malo više šuma, njen oblik bi se izgubio.[1]



a) vremenska serija bez šuma *b) vremenska serija sa šumom*

Slika 1.

Termin šum se često koristi u vezi sa podacima koji imaju prostornu ili vremensku komponentu. U takvim slučajevima, tehnike obrade signala ili slike se često mogu koristiti za smanjenje šuma i na taj način pomoći da se otkriju obrasci (signali) koji se mogu „izgubiti u šumu“. Bez obzira na to, eliminacija šuma je često teška, a veliki deo rada u predobradi podataka fokusira se na osmišljavanje robusnih algoritama koji daju prihvatljive rezultate čak i kada je šum prisutan.[2]

Podaci sa šumom (Noisy data) su skup podataka koji sadrži dodatne nepotrebne, lažne podatke. Skoro svi skupovi podataka sadrže određenu količinu neželjenog šuma. Ovakvi podaci se mogu filtrirati i objediniti u skup podataka boljeg kvaliteta. Termin *noisy data* se koristi i kao sinonim za podatke koje mašine ne mogu pravilno razumeti i protumačiti, kao što su nestrukturirani podaci. Podaci sa šumom nepotrebno povećavaju količinu potrebnog prostora za skladištenje i mogu negativno uticati na rezultate bilo koje analize podataka. [3]

Jedan od primera šuma u podacima je razgovor u prepunoj prostoriji. Ljudski mozak je odličan u filtriranju drugih razgovora tako da se čovek može fokusirati na jedan, ali ako je prostorija preglasna, postaje teško ili nemoguće pratiti razgovor koji sluša i gubi poruku koju pokušava da čuje. Na isti način, što se više dodatnih informacija dodaje skupu podataka, to postaje teže pronaći obrazac u podacima.[3]

1.1. Tipovi šuma u podacima

Veliki broj komponenti određuje kvalitet skupa podataka. Među njima, naziv klasa i vrednosti atributa direktno utiču na kvalitet skupa podataka za klasifikaciju. Kvalitet atributa odražava koliko dobro atributi karakterišu atribute za klasifikaciju, dok kvalitet naziva klasa ukazuje na to da li je klasa atributa ispravno dodeljena. Ako šum utiče na vrednosti atributa, ova sposobnost karakterizacije, a samim tim i kvalitet atributa, se smanjuje. Prilikom klasifikacije nazivi klasa se biraju na osnovu sledeće dve pretpostavke:[1]

1. Korelacija između atributa i klase: Pretpostavlja se da su atributi u određenoj meri povezani s klasom. Međutim, iako postoji korelacija, to ne znači da svi atributi imaju istu jačinu korelacije. Neki atributi mogu biti znatno jače povezani s klasom od drugih, a u takvim slučajevima ti atributi imaju veću važnost za klasifikaciju.
2. Slaba interakcija između atributa: Pretpostavlja se da su interakcije između atributa slabe [2], što znači da algoritmi učenja mogu zanemariti ove interakcije i razmatrati svaki atribut nezavisno prilikom kreiranja klasifikatora.

Na osnovu ova dva izvora informacija, u datom skupu podataka mogu se razlikovati dve vrste šuma[1]:

1. Klasni šum (takođe nazvan šum oznaka) Nastaje kada je primer nepravilno označen. Klasni šum može nastati iz nekoliko razloga, kao što su subjektivnost tokom procesa označavanja, greške pri unosu podataka ili neadekvatnost informacija korišćenih za označavanje svakog primera.

   Mogu se razlikovati dve vrste klasnog šuma:

   - Kontradiktorni primeri gde postoje duplikati primera u skupu podataka koji imaju različite oznake klasa.
   - Pogrešne klasifikacije (Misclassifications) gde su primeri označeni pogrešnim klasama.

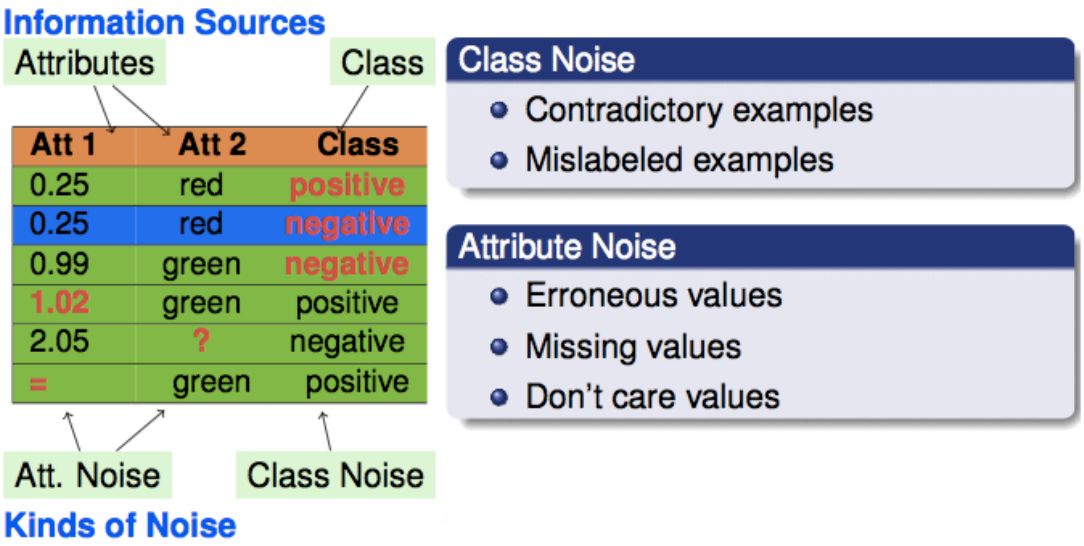
   Klasni šum obično se javlja na granicama klasa, gde primeri mogu imati slične karakteristike, iako se može pojaviti i u bilo kom drugom delu domena. Najčešći oblik klasnih grešaka u podacima iz stvarnog sveta su pogrešne klasifikacije.

2. Šum atributa se odnosi na korupciju u vrednostima jednog ili više atributa. Primeri šuma atributa su: pogrešne vrednosti atributa, nedostajuće ili nepoznate vrednosti atributa i nepotpuni atributi ili vrednosti „ne zanima me“.

   Šum atributa može poticati iz nekoliko izvora, kao što su ograničenja prenosa, kvarovi na senzorima, nepravilnosti u uzorkovanju i greške u transkripciji.

Pogrešne vrednosti atributa mogu biti potpuno nepredvidive, tj. nasumične, ili mogu implicirati malu varijaciju u odnosu na ispravnu vrednost.

Najčešći oblik atributnih grešaka su pogrešne vrednosti atributa. Šum atributa je štetniji od klasnih grešaka zbog svoje visoke korelacije sa oznakama klasa.



Slika 2. Klasni šum i šum atributa

Pored ova dva tipa šuma može se razlikovati i sistematski šum, nasumični šum i pozadinski šum.

Sistemski šum, takođe poznat kao pristrasnost, je vrsta šuma koji nastaje kada postoji konzistentna, predvidljiva greška ili pristrasnost u merenju ili prikupljanju podataka. Za razliku od nasumičnog šuma, koji se pojavljuje slučajno i može imati pozitivan ili negativan efekat na podatke, sistemski šum utiče na sve podatke na isti način, što može dovesti do pristrasnih rezultata i pogrešnih zaključaka.

Na primer, u slučaju sprovođenja testa inteligencije u zgradi koja je u procesu renoviranja, potencijalno bučna okolina će verovatno dovesti do nižih rezultata kod svih učesnika. Takvu vrstu sistematske greške bi trebalo izbegavati ili minimizirati, kako bi podaci precizno odražavali ono što se istražuje. U suprotnom, postoji rizik da rezultati ne budu validni za ono što je predmet istraživanja.[9]

Nasumična greška (koja se ponekad naziva i šum) nastaje zbog faktora koji nasumično utiču na merenje promenljive od interesa. Nasumičan šum je često veliki deo šuma u podacima.[5] Nasumičan šum u signalu meri se kao odnos signala i šuma (Signal-to-Noise Ratio). Nasumičan šum sadrži skoro jednake količine širokog spektra frekvencija i naziva se beli šum (kao što se boje svetlosti kombinuju da bi se dobila

bela). Nasumičan šum je neizbežan problem. Ona utiče na procese prikupljanja i pripreme podataka, gde se često javljaju greške.[5]

Na primer, nasumična greška može biti raspoloženje učesnika prilikom polaganja testa inteligencije, što može uticati na rezultate nekih učesnika, ali ne na sve. Nasumične greške nemaju sistematski uticaj na celokupan uzorak, zbog čega obično ne utiču na prosečne rezultate grupe.[9]

Pozadinski šum (background noise) odnosi se na neželjene ili nebitne informacije koje su prisutne u podacima, a koje mogu ometati ili otežati proces analize i donošenje zaključaka.[8] Ovaj šum može poticati iz različitih izvora, kao što su nesavršenosti u merenju, spoljašnji faktori ili slučajne greške u prikupljanju podataka.

U kontekstu analize podataka, pozadinski šum može prikriti važne signale ili obrasce u podacima, čineći ih teže uočljivim ili manje preciznim. Na primer, u audio snimcima, pozadinski šum može biti zvuk okoline koji otežava prepoznavanje glavnog zvuka ili govora. Ključni cilj u analizi podataka je smanjenje ili filtriranje pozadinskog šuma kako bi se izvukli jasniji i tačniji podaci koji će omogućiti bolje donošenje odluka i pouzdanije rezultate.

1.2. Razlika izmedju šuma i anomalije

I šum i anomalija su vrste nepravilnosti u podacima, ali služe različitim svrhama i imaju različite efekte na analizu podataka.

Šum se odnosi na nasumične fluktuacije ili greške u podacima koji ne nose nikakve značajne informacije. Šum je obično prisutan u svim podacima u određenoj meri i može biti uzrokovan različitim faktorima kao što su greške merenja, faktori okoline ili problemi sa prikupljanjem podataka. Smatra se da je šum deo prirodne varijabilnosti u podacima i obično je ravnomerno raspoređen po skupu podataka. Šum može uticati na tačnost statističkih analiza i modela mašinskog učenja uvođenjem nasumičnih varijacija koje mogu da prikriju obrasce u podacima. [10]

Anomalije su tačke podataka koje značajno odstupaju od ostatka podataka u skupu podataka. Izgledaju kao da ne pripadaju posmatranom skupu podataka. Anomalije nisu nasumične kao šum. One predstavljaju zapažanja koja se razlikuju od većine tačaka podataka. Anomalije mogu nastati zbog grešaka u merenju, eksperimentalne greške ili stvarnih anomalija u podacima. Anomalije mogu imati značajan uticaj na analizu podataka iskrivljavanjem statističkih mera kao što su srednja vrednost i standardna devijacija. Ponekad mogu sadržati vredne informacije ili ukazivati na važne obrasce u podacima, zbog čega je važno pažljivo razmotriti da li da ih treba ukloniti ili zadržiti u analizi. Anomalija nije lažna vrednost ili prazno značenje. Definitivno je i tačno, ali kada je povezano sa drugim torkama u modelu, jednostavno nije u istom opsegu.

Anomalija neće nužno dovesti do neuspeha modela, već samo do netačnog odgovora. S druge strane, šum će gotovo sigurno uzrokovati da model ne uspe, u 9 od 10 slučajeva.[11]

Ukratko, šum predstavlja nasumične greške u podacima koji ne nose smislene informacije, dok su anomalije tačke podataka koje značajno odstupaju od ostatka podataka i mogu imati značajan uticaj na analizu podataka. Šum je generalno ujednačena i sveprisutna, dok su anomalije različita i potencijalno informativna.

1.3. Izvor šuma u podacima

Razlike u stvarnim merenim podacima u odnosu na stvarne vrednosti proizilaze iz više faktora koji utiču na merenje. [4] Šum u podacima može nastati zbog kvarova na hardveru, grešaka u programiranju i ostalim nerelevantnim ili pogrešnim informacijama koje unose programi za prepoznavanje govora ili optičkog karaktera. Pravopisne greške, industrijske skraćenice i sleng takođe mogu da ometaju mašinsko čitanje. Prirodne fluktuacije u senzorima i merenjima mogu dodati dodatni šum očitavanjima. Prikupljanje previše širokog skupa podataka takođe može otežati analizu.[3]

Šum ima dva glavna izvora greške nastale zbog rada mernih alata i nasumične greške nastale od strane stručnjaka prilikom prikupljanja ili obrade podataka.[6]

Nepravilno filtriranje može dodati šum ako se filtrirani signal tretira kao direktno mereni signal. Na primer, digitalni filteri tipa konvolucije, kao što je pokretni prosek, mogu imati nuspojave kao što su kašnjenja ili skraćivanje vrhova. Digitalni filteri za diferenciranje pojačavaju nasumičan šumu originalnim podacima.

Ljudi ponekad namerno menjaju podatke kako bi postigli željene rezultate. Podaci koji izgledaju uredno i bez mnogo odstupanja mogu predstavljati onoga ko ih prikuplja u boljem svetlu, pa može postojati motivacija da se uklone ili izmene anomalije u podacima kako bi izgledali bolje nego što zaista jesu.

1.4. Uticaj šuma na kvalitet podataka

Šum u podacima može značajno uticati na analizu podataka i procese mašinskog učenja. On može iskriviti osnovne obrasce, uvesti pristrasnosti u algoritme i smanjiti tačnost i pouzdanost prediktivnih modela. Neki od glavnih uticaja šuma na analizu podataka su:

1. **Smanjena tačnost**: Šum može uneti nasumične varijacije koje odstupaju od stvarnih obrazaca u podacima, što dovodi do grešaka prilikom klasifikacije, regresije ili grupisanja.

2. **Overfitting**: Zbog šuma, modeli mašinskog učenja mogu postati previše prilagođeni šumu umesto da pronađu stvarne odnose u podacima. To može rezultirati modelima koji dobro rade na podacima za obuku, ali se loše generalizuju na nove podatke.
3. **Povećana nesigurnost**: Šum može povećati nesigurnost i različitost u rezultatima analize podataka, što otežava donošenje pouzdanih odluka ili predviđanja na osnovu tih podataka.

1.5. Uticaj šuma na kvalitet klasifikacije

U klasifikaciji, šum može negativno uticati na performanse sistema u smislu tačnosti klasifikacije, vreme izgradnje, veličina i interpretabilnost izgrađenog klasifikatora [1]. Prisustvo suma u podacima može uticati na suštinske karakteristike klasifikacije. Sum može stvoriti male grupe instanci određene klase u delovima prostora koji odgovara drugoj klasi, ukloni instance koje se nalaze u ključnim oblastima unutar konkretne klase ili narušiti granice klasa i povećati prklapanje među njima. Ove izmene kvare znanje koje se može izvući od problema i pokvariti klasifikatore izgrađene od tih podataka sa šumom s poštovanjem do originalnih klasifikatora izgrađenih od čistih podataka koji predstavljaju najtačnije implicitno poznavanje problema [1].

Šum je posebno relevantan u nadgledanim problemima, gde menja odnos između informativnih karakteristika i rezultata mere. Zbog toga je šum posebno proučavan u klasifikaciji i regresiji gde šum ometa izvlačenje znanja iz podataka i kvari modele dobijene korišćenjem tih podataka sa šumom kada se uporede sa modelima naučenim iz čistih podataka iz istog problema, koji predstavljaju pravi implicitno poznavanje problema []. U tom smislu, robusnost [] je sposobnost algoritma da izgradi modele koji su neosetljivi na oštećenja podataka i manje pate od uticaja šuma; to jest, što je algoritam robusniji, to su sličniji modeli izgrađeni od čistih i podataka sa šumom.[1]

2. Metodi za uklanjanje šuma

Uklanjanje šuma iz podataka je ključni korak u osiguravanju tačnosti i pouzdanosti analize podataka i modela mašinskog učenja. Šum može značajno narušiti kvalitet podataka, što može dovesti do pogrešnih zaključaka, prekomernog prilagođavanja modela i povećane nesigurnosti u rezultatima. Da bi se postigli precizniji i robusniji modeli, neophodno je primeniti odgovarajuće metode za identifikaciju i uklanjanje šuma. Ovaj podnaslov će istražiti različite tehnike i pristupe koji se koriste za smanjenje šuma u podacima, uključujući metode prečišćavanja podataka, filtriranja i napredne algoritme koji omogućavaju bolju detekciju i eliminaciju neželjenih varijacija. Razumevanje i primena ovih metoda ključni su za poboljšanje kvaliteta podataka i osiguravanje validnih analitičkih rezultata.

2.1. Tradicionalne statističke metode

Tradicionalne statističke metode igraju ključnu ulogu u identifikaciji i uklanjanju šuma iz podataka. Ove metode se oslanjaju na matematičke principe za prepoznavanje anomalija i varijacija koje odstupaju od očekivanih obrazaca u podacima. Kroz upotrebu srednje vrednosti, medijane, standardne devijacije i drugih osnovnih statističkih mera, moguće je precizno detektovati anomalije koje predstavljaju šum i ukloniti ih kako bi se obezbedila veća tačnost i pouzdanost rezultata. Tradicionalne statističke metode takođe omogućavaju primenu različitih tehnika za transformaciju podataka, poput normalizacije i standardizacije, koje pomažu u smanjenju efekata šuma i poboljšanju performansi analitičkih modela.

Tradicionalne statističke metode za uklanjanje šuma često se primenjuju na različite vrste podataka, uključujući:

1. **Numerički podaci**: Najčešće se sreću u analizi finansijskih izveštaja, istraživanjima i eksperimenalnim podacima gde se koriste osnovne statističke mere kao što su srednja vrednost, medijana i standardna devijacija za identifikaciju i uklanjanje outliera (neobičnih vrednosti) koje mogu predstavljati šum.
2. **Kvantitativni istraživački podaci**: U društvenim i prirodnim naukama, gde se tradicionalne metode koriste za procenu pouzdanosti merenja i za eliminaciju nepreciznih ili ekstremnih vrednosti koje mogu narušiti analize.
3. **Medicinski podaci**: U analizi medicinskih rezultata, kao što su rezultati laboratorijskih testova, gde se koriste statističke metode za prepoznavanje i uklanjanje netačnih merenja koja mogu biti uzrokovana greškama u instrumentima ili prikupljanju podataka.

4. **Eksperimentalni podaci**: U naučnim eksperimentima, gde se tradicionalne metode koriste za filtriranje netačnih rezultata koji mogu nastati zbog grešaka u eksperimentalnim procedurama ili merenju.
5. **Podaci u istraživanjima tržišta**: U analizi podataka o potrošačima, gde se koristi statistička analiza za uklanjanje neobičnih ili ekstremnih vrednosti koje mogu biti rezultat grešaka u unosu podataka ili neslaganja u odgovorima.
6. **Podaci o performansama**: U analizi performansi sistema ili uređaja, gde se koristi statistička analiza za uklanjanje podataka koji odstupaju od uobičajenih performansi zbog grešaka u merenju ili drugih faktora šuma.

Tradicionalne metode su često efikasne za prepoznavanje i uklanjanje šuma u ovim kontekstima, ali u slučajevima gde je šum kompleksniji ili ima specifične karakteristike, dodatne tehnike kao što su metode obrade signala mogu biti potrebne za postizanje boljih rezultata.

2.1.1. Srednja vrednost (Moving Average)

Srednja vrednost (Moving Average) je jedna od najjednostavnijih i najčešće korišćenih tehnika za uklanjanje šuma iz vremenskih serija podataka. Ideja ove metode je da se izračuna prosečna vrednost za određeni broj uzastopnih tačaka u podacima i ta vrednost se koristi kao novi podatak u seriji. Postoji nekoliko varijanti ove metode:

- Jednostavna pokretna sredina (Simple Moving Average - SMA): Ovo je osnovna varijanta, gde se izračunava aritmetička sredina za prethodni fiksni broj tačaka.
- Eksponencijalna pokretna sredina (Exponential Moving Average - EMA): Ovde se veća težina daje novijim podacima, čime se bolje prati trenutni trend u podacima.

Primena pokretne sredine može značajno smanjiti uticaj slučajnog šuma, ali može takođe rezultirati gubitkom detalja u podacima, posebno kada se koristi sa velikim prozorima za izračunavanje srednje vrednosti.

Srednja vrednost se često koristi za uklanjanje šuma kod podataka koji su numerički i gde je cilj smanjenje efekta nasumičnih varijacija.

Korišćenjem srednje vrednosti, moguće je značajno poboljšati kvalitet podataka i povećati tačnost analiza, posebno kada su prisutne nasumične varijacije koje ne predstavljaju pravu informaciju, već šum.

Dobre strane:

- **Jednostavnost**: Metoda je jednostavna za implementaciju i razumevanje.
- **Efikasnost**: Brzo izračunavanje, posebno kod velikih skupova podataka.

- **Smanjenje šuma**: Efikasno smanjuje nasumični šum i pomaže u izgladjivanju podataka.

Loše strane:

- **Izglađivanje korisnih podataka**: Može dovesti do gubitka važnih signala ili informacija, posebno u slučajevima kada postoji značajna promena u podacima.
- **Neefikasno kod impulsnog šuma**: Nije efikasno za uklanjanje šuma koji se pojavljuje u obliku ekstremnih vrednosti (impulsni šum).
- **Nije prilagođeno za podatke sa ivicama**: U obradi slika ili vremenskih serija, može zamagliti oštre ivice ili diskontinuitete.

2.1.2. Medijan filter

Medijan filter je popularna metoda za uklanjanje šuma, naročito efikasan kod impulsnog šuma, koji se manifestuje kao nagle i velike oscilacije u podacima. Za razliku od aritmetičke sredine, koja može biti podložna uticaju ekstremnih vrednosti, medijan tačnije odražava centar raspodele podataka unutar prozora. Na primer, niz podataka [2, 3, 100, 3, 2], medijan će biti 3, dok bi prosečna vrednost bila znatno viša zbog uticaja ekstremne vrednosti 100.

Ova metoda se često koristi u obradi slike za uklanjanje soli i biber šuma, gde pojedinačni pikseli postaju ekstremno crni ili beli. Medijan filter je posebno koristan kada je potrebno očuvati strukturne detalje u podacima, kao što su ivice u slikama ili važni pikovi u vremenskim serijama, dok se uklanjaju impulzivni ili ekstremni oblici šuma.

Dobre strane:

- **Očuvanje ivica**: Efikasno uklanja šum dok zadržava ivice i strukturne karakteristike podataka, što je posebno korisno u obradi slika.
- **Robustnost na impulsni šum**: Veoma efikasan za uklanjanje impulsnog šuma ("sol i biber" šuma).
- **Fleksibilnost**: Može se primeniti na različite tipove podataka, uključujući slike, audio, i vremenske serije.

Loše strane:

- **Složenost**: Može biti računarski zahtevnije od jednostavnih metoda kao što je srednja vrednost, posebno za velike podatke.
- **Potencijalno zamućenje podataka**: Iako je efikasan za uklanjanje impulsnog šuma, medijan filter može uzrokovati blago zamućenje na mestima gde nema šuma.

- **Neefikasnost za kontinuirane promene**: Nije idealan za podatke sa glatkim, kontinuiranim promenama jer može narušiti tok signala.

2.1.3. Kalmanov filter

Kalmanov filter je napredna metoda za procenu stanja dinamičkih sistema, kao što su sistemi sa vremenskim promenama. Ova metoda koristi iterativni proces koji kombinuje prethodne procene sa novim merenjima kako bi smanjila šum. Kalmanov filter se koristi u raznim aplikacijama, od navigacionih sistema do praćenja objekata u stvarnom vremenu.

Kalmanovo filtriranje koristi pristup rekurzivne Bajesove procene za filtriranje šuma iz vremenskih serija podataka u dinamičkim sistemima. Radi na sledeći način:

1. Pravi se početna pretpostavka o stanju sistema
2. Meri se šum i rafinira procenu
3. Ponavlja se ovaj ciklus predikcije i ažuriranja za filtriranje šuma

Njegova snaga leži u sposobnosti da precizno predviđa i koriguje buduće stanja sistema, čak i u prisustvu neizvesnosti i šuma. Kalmanovo filtriranje je optimalno za Gausove distribucije šuma. Omogućava praćenje i predviđanje ponašanja sistema u realnom vremenu u prisustvu šuma. Primeri upotrebe uključuju GPS navigaciju, ekonomsko predviđanje i praćenje objekata.[12]

Kalmanov filter je efikasan u situacijama gde je potrebno kontinuirano pratiti i predviđati stanje sistema u prisustvu šuma i nesigurnosti u merenjima, čineći ga jednim od ključnih alata u oblastima kao što su navigacija, robotika, finansije i medicina. Kalmanov filter se često koristi u automobilskoj industriji, avijaciji, robotici i mnogim drugim oblastima gde je potrebna tačna procena u stvarnom vremenu.

Dobre strane:

- **Preciznost**: Visoko precizan za praćenje i predikciju stanja dinamičkih sistema.
- **Realno vreme**: Efikasan za obradu podataka u realnom vremenu, sa sposobnošću ažuriranja procena sa svakim novim merenjem.
- **Kombinovanje informacija**: Efikasno kombinuje informacije iz modela i merenja, smanjujući uticaj šuma.
- **Primena u različitim oblastima**: Široko primenljiv u navigaciji, robotici, finansijama, i biomedicini.

Loše strane:

- **Složenost**: Relativno složen za implementaciju, zahteva dobro poznavanje statistike i matematičkog modelovanja.

- **Zavisnost od modela**: Efikasnost zavisi od kvaliteta modela dinamike sistema; ako je model netačan, procene mogu biti loše.
- **Osetljivost na ne-gausovski šum**: Kalmanov filter podrazumeva da je šum gausovskog tipa; ako šum nije gausovski, performanse filtra mogu biti narušene.
- **Potreba za inicijalnim uslovima**: Uspeh Kalmanovog filtera zavisi od pravilne postavke početnih uslova, što može biti izazovno u nekim aplikacijama.

2.2. Metode zasnovane na obradi signala

Metode zasnovane na obradi signala koriste se za uklanjanje šuma iz podataka, posebno u slučajevima kada su podaci vremenski ili prostorno kontinuirani, kao što su audio snimci, slike, ili senzorski podaci. Ove metode koriste različite tehnike za filtriranje neželjenih komponenti iz signala, zadržavajući pritom ključne informacije.

2.2.1. Fourierova transformacija

Fourierova transformacija je matematička tehnika koja se koristi za transformisanje signala iz vremenskog ili prostornog domena u frekvencijski domen. U kontekstu uklanjanja šuma, može pomoći u identifikaciji i filtriranju šuma predstavljajući signal kao kombinaciju različitih frekvencija. Relevantne frekvencije mogu se zadržati, dok se frekvencije šuma mogu filtrirati.[13] Time omogućava identifikaciju dominantnih frekvencija u signalu, što pomaže u izolaciji i uklanjanju šuma visokih frekvencija.

Fourierova transformacija je naročito korisna kada šum ima poznatu frekvencijsku komponentu, koja može biti filtrirana pomoću niskopropusnih filtera.

Fourierova transformacija pretpostavlja da je signal stacionaran, što znači da se njegove statističke karakteristike ne menjaju tokom vremena. U stvarnim podacima, ovaj uslov nije uvek zadovoljen, zbog čega se koriste naprednije tehnike kao što su vremensko-frekvencijski pristupi.

Fourierova transformacija se koristi za uklanjanje šuma kod podataka koji se mogu predstaviti kao vremenski ili prostorni signali, kao što su audio zapisi, digitalne slike, telekomunikacioni signali, podaci iz senzora, i meteorološki podaci. Ova metoda je posebno efikasna za identifikaciju i filtriranje neželjenih frekvencijskih komponenti, što omogućava zadržavanje korisnih informacija dok se uklanja šum, čime se poboljšava kvalitet i tačnost analize signala.

Fourierova transformacija je vrlo efikasna za analizu periodičnih signala i identifikaciju frekvencijskih komponenti. Omogućava lako filtriranje frekvencija koje predstavljaju šum, dok se očuvavaju korisni delovi signala. Ova metoda je široko primenjiva u analizi

zvuka, slika i drugih vrsta signala, a njeno razumevanje i implementacija su dobro dokumentovani.

Fourierova transformacija može biti manje efikasna za nestacionarne signale jer ne pruža informacije o vremenskim promenama frekvencija. Može se suočiti sa izazovima u analizi signala sa promenjivim karakteristikama i u situacijama gde je šum neuniforman ili varijabilan u vremenu.

2.2.2. Wavelet transformacija

Wavelet transformacija je metoda koja omogućava analizu signala u vremensko-frekvencijskom domenu. Za razliku od Fourierove transformacije, wavelet transformacija koristi kratke oscilacije (wavelete) da analizira signal na različitim vremenskim skalama. Ova metoda je naročito korisna za signale sa promenljivim karakteristikama, jer omogućava bolju lokalizaciju šuma u vremenu i frekvenciji.

Ovaj pristup uklanja visokofrekventni šum dok zadržava originalne karakteristike signala. Wavelet denoising je koristan za nestacionarne signale i može se prilagoditi različitim vrstama šuma. Široko se koristi za obradu slika, ali je takođe primenjiv za vremenske serije podataka i obradu signala.[12]

Wavelet denoising, ili uklanjanje šuma pomoću wavelet transformacije, često se koristi u analizi vremenskih serija, slikama i bioinformatici. Postupak uključuje razlaganje signala u wavelete, zatim primenu praga za eliminaciju malih koeficijenata (koji obično predstavljaju šum), i na kraju rekonstrukciju signala. Ova tehnika je efikasna jer zadržava bitne karakteristike signala dok eliminiše šum.

Wavelet transformacija je izuzetno korisna za analizu nestacionarnih signala jer omogućava razdvajanje signala u vremensko-frekvencijskom domenu. Ova metoda je fleksibilna i može se prilagoditi različitim vrstama podataka, što je čini pogodnom za složene i varijabilne signale. Omogućava detaljnu analizu i filtriranje signala na različitim skalama, što je korisno za precizno uklanjanje šuma.

Implementacija i razumevanje wavelet transformacije mogu biti kompleksni, posebno zbog potrebe za odabirom odgovarajuće wavelet funkcije. Takođe, ova metoda može biti računarski zahtevna, što može predstavljati izazov za obradu velikih skupova podataka ili u aplikacijama u realnom vremenu.

2.2.3. Savitzky-Golay filter

Savitzky-Golay filter je metoda za izravnavanje podataka koja koristi polinomnu aproksimaciju kroz prozore podataka. Ova metoda se koristi za očuvanje oblika signala, posebno u situacijama gde je potrebno zadržati pikove i nagle promene u signalu.

Za razliku od drugih filtera, koji mogu značajno smanjiti varijabilnost u podacima, Savitzky-Golay filter omogućava očuvanje osnovne strukture signala. Ovo je posebno korisno u spektralnoj analizi, gde je važno zadržati karakteristične pikove. Ova metoda se koristi u raznim naučnim disciplinama, kao što su hemija, biologija i obrada signala.

Savitzky-Golay filter je dizajniran da minimizira gubitak podataka, zadržavajući originalne oblike signala kao što su vrhovi i padovi. Filter radi tako što aproksimira lokalne segmente signala polinomom i koristi ga za glatko filtriranje, što omogućava zadržavanje informacija bolje nego obični klizni prosek.

Savitzky-Golay filter je idealan za primene gde je potrebno ukloniti šum, ali bez značajnog narušavanja strukture i oblika originalnog signala. Ovaj filter je efikasan u zadržavanju originalnih struktura signala, što ga čini pogodnim za primene gde je očuvanje detalja ključno. Takođe, njegova implementacija je relativno jednostavna u poređenju sa drugim metodama.

Savitzky-Golay filter može biti manje efikasan kada se suočava sa jakim šumovima, a kod viših polinomijalnih stepeni može doći do preklapanja podataka. Iako je dobar u očuvanju strukture signala, možda nije idealan za sve vrste šuma i može imati ograničenja u određivanju optimalnih parametara za specifične primene.

2.3. Metode mašinskog učenja

Metode mašinskog učenja nude sofisticirane pristupe za rešavanje problema šuma u podacima. Ove metode koriste algoritme i modele koji mogu automatski prepoznati i filtrirati šum, čime se poboljšava kvalitet podataka i omogućava tačnija analiza. Od robustnih regresionih tehnika do naprednih algoritama za detekciju outliera, mašinsko učenje nudi širok spektar alata za adresiranje različitih vrsta šuma.

2.3.1. Principal component analysis (PCA)

Principal component analysis (PCA) je tehnika redukcije dimenzionalnosti koja se koristi za identifikaciju najvažnijih varijacija u podacima. PCA projektuje podatke na nove ose, gde su prve komponente one sa najvećom varijansom. Ove nove varijable se nazivaju “glavne komponente.” Ostatak komponenti, koje često predstavljaju šum, može

biti odbačen, čime se smanjuje uticaj šuma na podatke. Cilj PCA-a je da ukloni šum iz signala ili slike dok čuva osnovne karakteristike. Ovaj pristup je i geometrijski i statistički. PCA projektuje ulazne podatke duž različitih osa, smanjujući dimenziju ulaznog signala ili podataka. Ovaj fenomen se takođe naziva “redukcija dimenzionalnosti.”[14]

Projekcijom podataka na smanjen skup glavnih komponenti, PCA može pomoći u smanjenju šuma fokusiranjem na najinformativnije dimenzije podataka, dok odbacuje dimenzije povezane sa šumom.[13]

PCA se često koristi u obradi slike, genetici, i ekonomiji za uklanjanje šuma iz podataka sa više varijabli. Na primer, u obradi slike, PCA može biti korišćena za smanjenje šuma eliminacijom komponenti sa niskom varijansom, koje obično predstavljaju šum.

Prednosti:

1. **Smanjenje dimenzionalnosti**: PCA (Principal Component Analysis) može značajno smanjiti dimenzionalnost podataka, što može pomoći u uklanjanju šuma, posebno u visokodimenzionalnim skupovima podataka. Ovo smanjenje često uklanja nebitne varijable koje mogu sadržavati šum.
2. **Očuvanje ključnih informacija**: PCA identifikuje glavne komponente koje obuhvataju najveći deo varijanse u podacima. Time pomaže da se očuvaju ključne informacije i obrasci u podacima, dok se smanjuje uticaj šuma koji je često prisutan u manjim komponentama.
3. **Vizualizacija podataka**: Smanjivanje dimenzionalnosti uz pomoć PCA može omogućiti lakšu vizualizaciju i interpretaciju podataka, što može pomoći u identifikaciji i analizi šuma.

Mane:

1. **Gubitak interpretabilnosti**: PCA može dovesti do gubitka interpretabilnosti, jer glavne komponente nisu uvek direktno povezane sa originalnim atributima. Ovo može otežati razumevanje koji tačno atributi utiču na specifične komponente.
2. **Neefikasnost za nelinearne šumove**: PCA je linearna tehnika i može biti manje efikasna u uklanjanju nelinearnih šumova ili kompleksnih obrazaca u podacima. U slučajevima kada je šum nelinearan, PCA možda neće pružiti optimalne rezultate.
3. **Zavisnost od skupa podataka**: Efikasnost PCA zavisi od skupa podataka koji se koristi za analizu. Ako su podaci sa šumom već prisutni u velikoj meri, PCA može samo delimično ukloniti šum i možda neće biti dovoljno efikasna u potpunosti ga eliminisati.

2.3.2. Autoenkoderi

Autoenkoderi su specijalne vrste neuronskih mreža koje se sastoje od enkodera i dekodera. Enkoder kompresuje ulazne podatke u reprezentaciju sa nižim dimenzijama, dok dekoder rekonstruiše originalne podatke iz te reprezentacije.

Kada su trenirani na podacima bez šuma, autoenkoderi mogu prilično dobro naučiti da razlikuju signal od šuma. Jedan od načina na koji se to postiže je dodavanjem šuma originalnim podacima tokom treninga, a zatim treniranjem mreže da rekonstruiše čiste podatke.

Mogu se koristiti kao uređaji za uklanjanje šuma pružanjem podataka sa šumom kao ulaza i dobijanjem čistih podataka kao izlaza. Glavna ideja iza autoenkodera za uklanjanje šuma je da se prisili skriveni sloj da nauči dodatne robusne karakteristike. Zatim se obučava autoenkoder da rekonstruiše ulazne podatke iz degradirane verzije minimizovanjem gubitka. [14]

Koristi se za kompresiju slika, denoising slike, kao i za predikciju vremenskih serija. Autoenkoderi su fleksibilni i mogu se prilagoditi specifičnim karakteristikama podataka, što ih čini dobrim u uklanjanju šuma.

Prednosti:

1. **Nelinearna obrada**: Autoenkoderi mogu modelirati kompleksne, nelinearne odnose između ulaznih podataka, što ih čini efikasnim za uklanjanje nelinearnih šumova. Ova sposobnost omogućava bolje očuvanje značajnih karakteristika podataka dok se uklanja šum.
2. **Učenje značajnih karakteristika**: Autoenkoderi uče niske dimenzionalne reprezentacije podataka, čime automatski uče relevantne karakteristike i obrasce, a ne samo površne informacije. Ovo omogućava bolje filtriranje šuma jer autoenkoderi prepoznaju i očuvavaju važne informacije.
3. **Prilagodljivost**: Autoenkoderi su fleksibilni i mogu se prilagoditi različitim vrstama podataka i šuma. Mogu se koristiti za obrada slika, zvuka, i drugih tipova podataka, čineći ih široko primenljivim u različitim domenima.

Mane:

1. **Zahtevnost u obuci**: Obuka autoenkodera može biti zahtevna i vremenski intenzivna. Za postizanje dobrih rezultata potrebni su kvalitetni podaci za obuku, kao i pažljivo podešavanje hiperparametara modela.
2. **Overfitting**: Ako autoenkoder nije pravilno regulisan, može doći do overfittinga, gde model uči šum u podacima umesto da ga eliminiše. Ovo može rezultirati lošim performansama na neviđenim podacima ili smanjenjem generalizacije.
3. **Kompleksnost modela**: Autoenkoderi mogu biti kompleksni za implementaciju i optimizaciju, posebno kada se koriste napredne arhitekture kao što su

konvolucioni ili rekurentni autoenkoderi. Takođe, može biti potrebno puno resursa za obuku, što može biti izazov u okruženjima sa ograničenim računarstvom.

2.3.3. GAN (Generative Adversarial Networks)

Generative Adversarial Networks (GANs) su široko rasprostranjene u preprocesiranju fotografija a sada se primenjuju i u drugim oblastima preprocesiranja podataka. GAN je napredna tehnika mašinskog učenja koja se sastoji od dva podmodela generatora i diskriminatora. Generator je primarni model u toku treniranja GAN mreže, njegova uloga je da predvidi podatke, u ovom slučaju da generiše podatke koje ne sadrže šum na osnovu ulaznih podataka sa šumom. Deskriminator ima ulogu da razlikuje stvarne podatke od lažnih u cilju poboljšanja tačnosti generatora.

GANs se koriste za uklanjanje šuma tako što se treniraju da generišu "čiste" podatke, pri čemu diskriminator služi kao kontrola kvaliteta. Na primer, u obradi slike, GAN može biti obučen da rekonstruiše slike bez šuma, čak i kada je šum kompleksan i nije lako definisati pravila za njegovo uklanjanje.

GAN-ovi su posebno korisni u situacijama gde je šum vrlo nepravilnog oblika ili gde su tradicionalne metode neuspešne. GAN-ovi zahtevaju dosta vremena za treniranje i skloni su problemima sa stabilnošću, ali njihova fleksibilnost i sposobnost da generišu visokokvalitetne podatke ih čini jednom od najboljih metoda u modernoj obradi podataka.

3. Praktična primena

Za potrebe demonstracije svih metoda korišćeno je tri seta podataka:

- Set 1 za demonstriranje statističkih metoda
- Set 2 za demonstriranje metoda mašinskog učenja
- Set 3 za demonstriranje GAN metode

3.1. Statističke metode

Ideja praktičnog dela ovog rada je prikazivanje kako šum utiče na klasifikaciju i kako se određenim metodama može poboljšati rezultat klasifikacije.

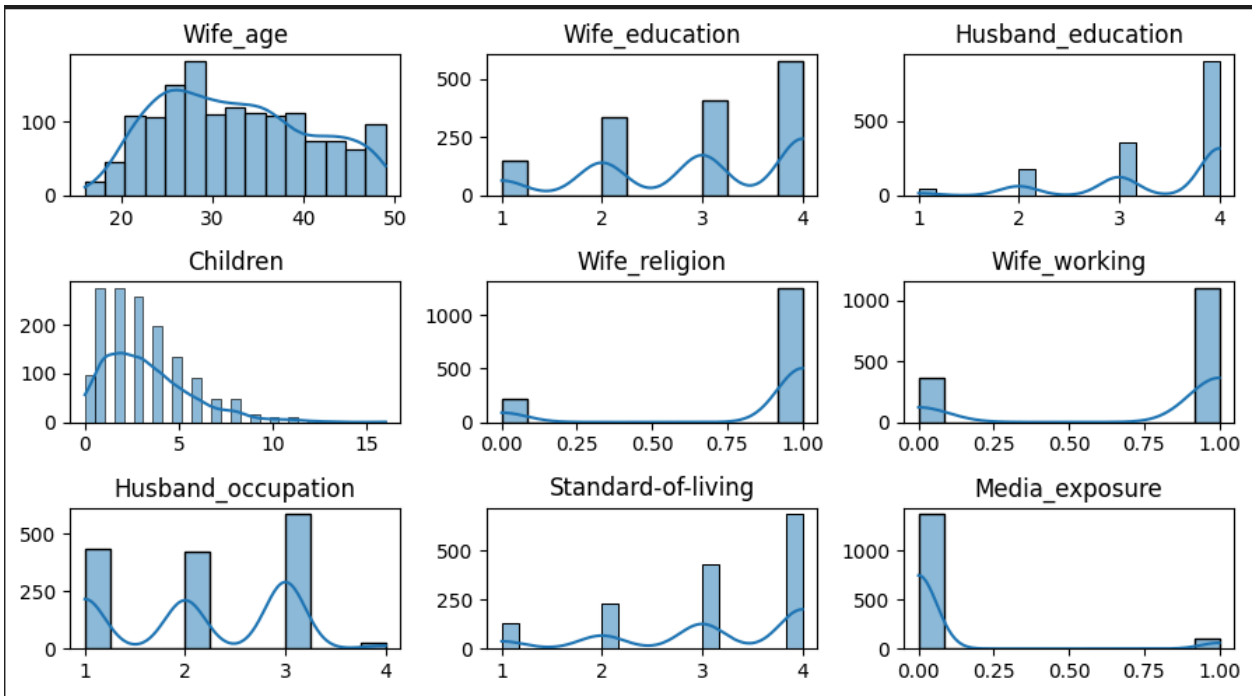
Prvi set podataka, korišćen za demonstriranje statističkih metoda, je podskup Nacionalnog istraživanja prevalencije korišćenja kontraceptiva iz Indonezije iz 1987. godine. Uzorci su prikupljeni od udatih žena koje u trenutku intervjua nisu bile trudne ili nisu bile sigurne u svoje stanje.

Ovaj dataset sadrži sledeće kolone:

1. Wife_age: Starost supruge, izražena u godinama.
2. Wife_education: Nivo obrazovanja supruge, kategorizovan na sledeći način:
    - 1: Nema obrazovanje
    - 2: Osnovno obrazovanje
    - 3: Srednje obrazovanje
    - 4: Visoko obrazovanje
3. Husband_education: Nivo obrazovanja supruga, koristeći iste kategorije kao kod supruge.
4. Children: Broj dece koju par ima, kao celobrojna vrednost.
5. Wife_religion: Religijska pripadnost supruge, predstavljena kao binarna vrednost:
    - 0: Ne-islam
    - 1: Islam
6. Wife_working: Informacija o tome da li je supruga zaposlena:
    - 0: Ne radi
    - 1: Radi
7. Husband_occupation: Zanimanje supruga, kategorizovano na sledeći način:
    - 1: Poljoprivreda
    - 2: Radnik u industriji
    - 3: Usluge
    - 4: Profesionalna zanimanja
8. Standard-of-living: Indeks životnog standarda porodice, klasifikovan kao:
    - 1: Nizak

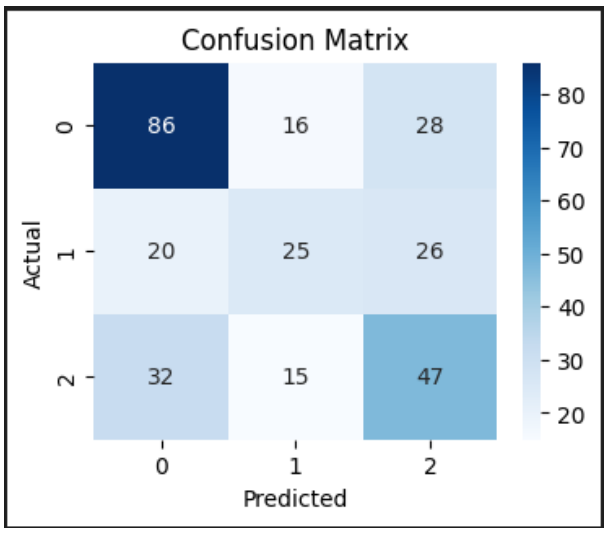
- 2: Srednji
- 3: Visok
- 4: Veoma visok

9. Media_exposure: Izloženost porodice medijima:
   - 0: Nema izloženost
   - 1: Izloženi
10. Contraceptive_method: Ciljna promenljiva koja predstavlja trenutni izbor metode kontracepcije kod žene, klasifikovana kao:
    - 1: Nema korišćenja kontracepcije
    - 2: Kratkoročne metode (npr. pilule, kondomi)
    - 3: Dugoročne metode (npr. IUD, sterilizacija)

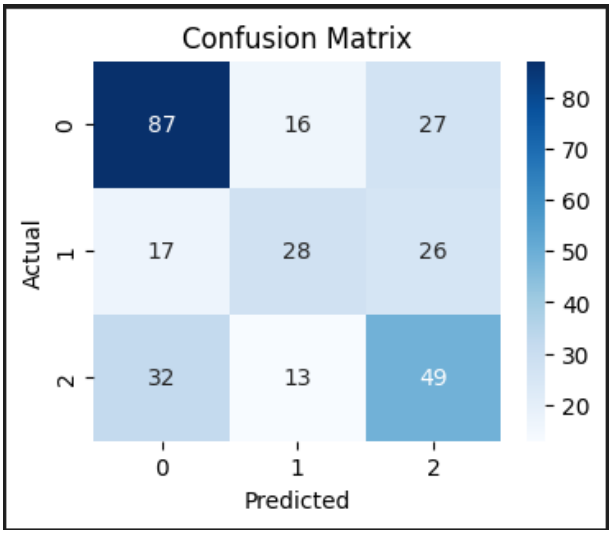


Slika 3. Raspodela kolona

Korišćena su dva modela klasifikacije KNN klasifikator i RandomForest klasifikator. Na Slici 4 i Slici 5 prikazani su rezultati primene klasifikatora pre primene metoda za redukciju šuma.



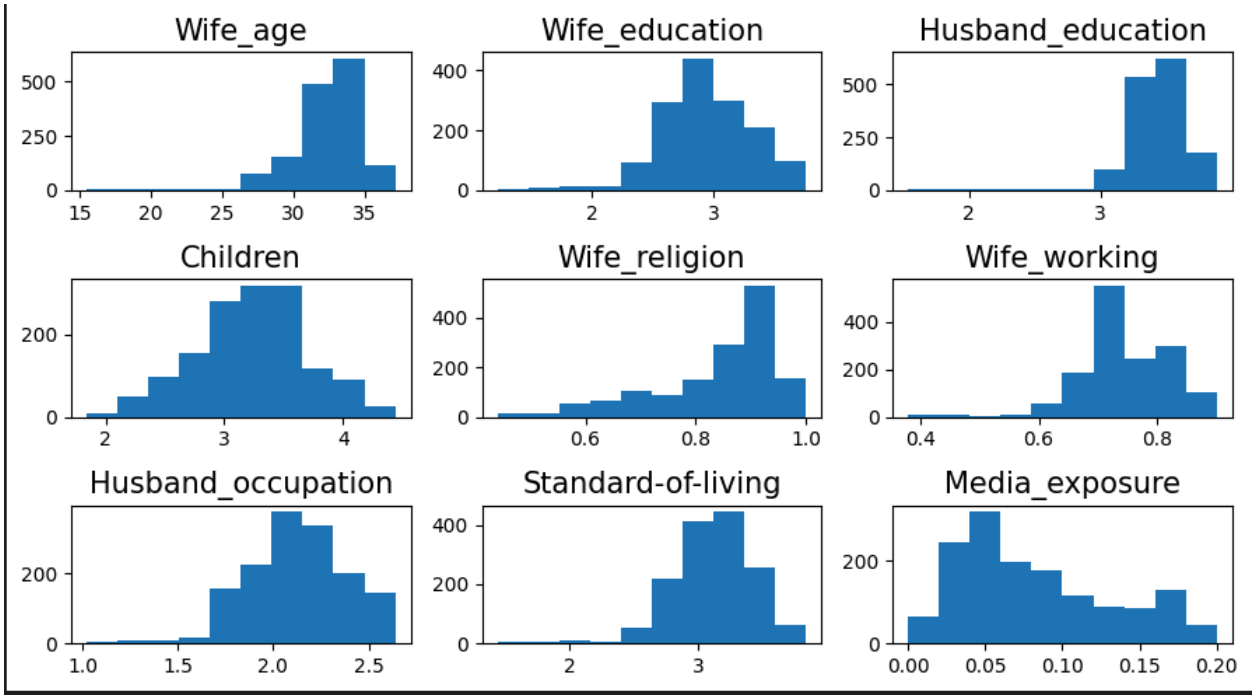
Slika 4. Random Forest klasifikator



Slika 5. KNN klasifikator

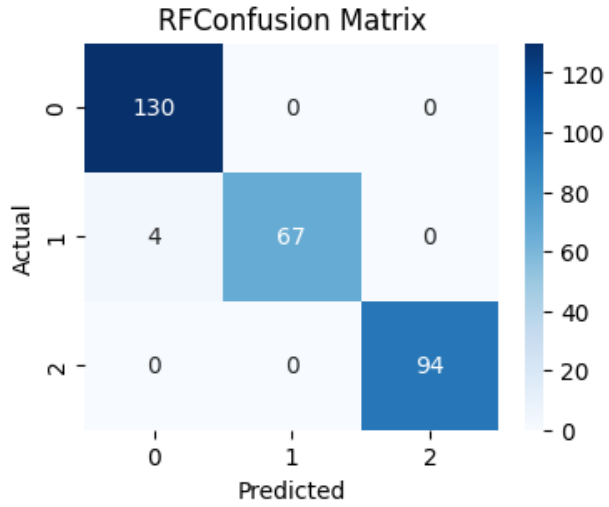
3.1.2. Metod srednje vrednosti

Raspodela kolona nakon primene metode srednje vrednosti:

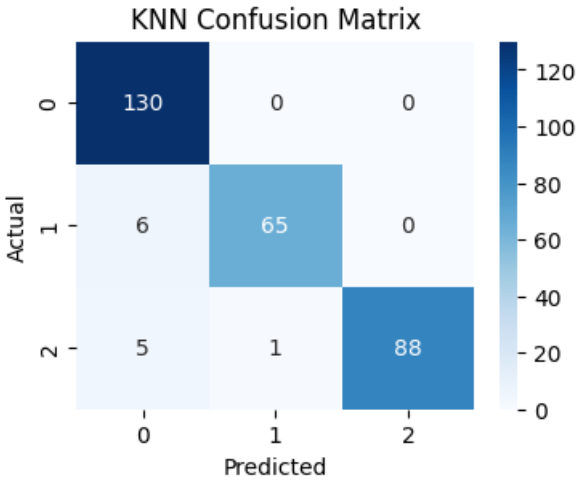


Slika 6. Raspodela kolona nakon primene metode srednje vrednosti

Rezultati klasifikacije:

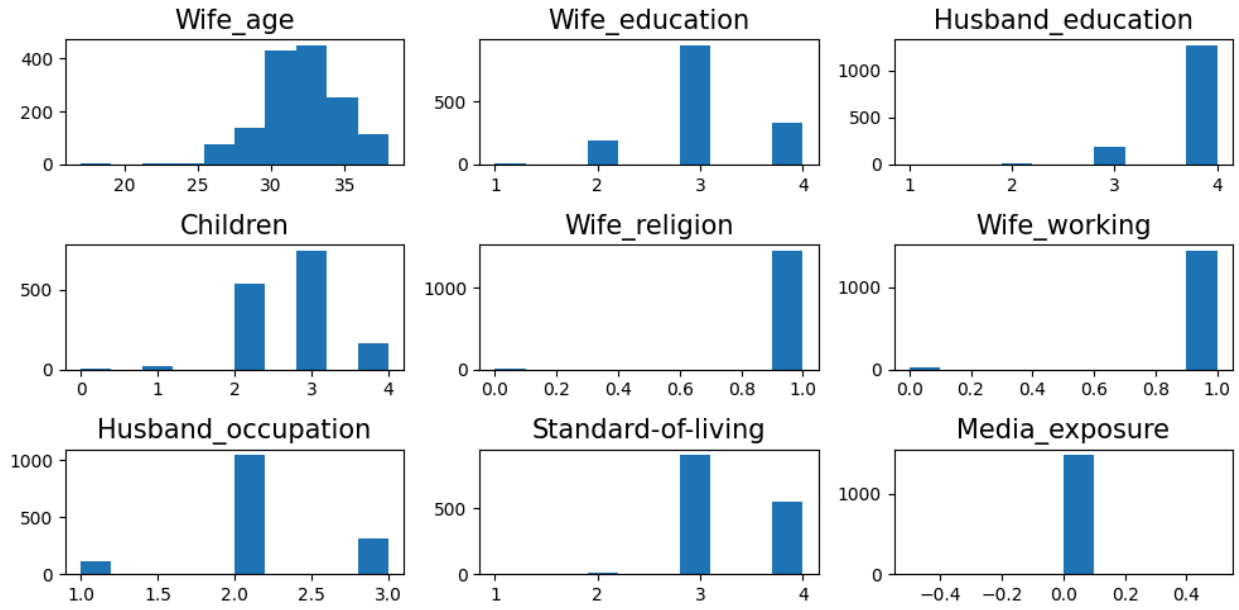


Slika 7. Random forest nakon primene metode srednje vrednosti

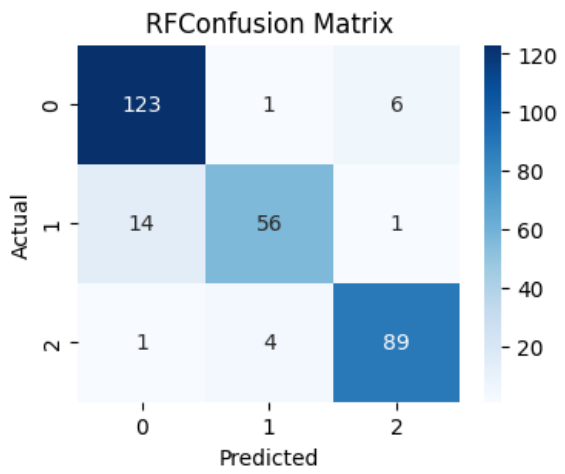


Slika 8. KNN nakon primene metode srednje vrednosti

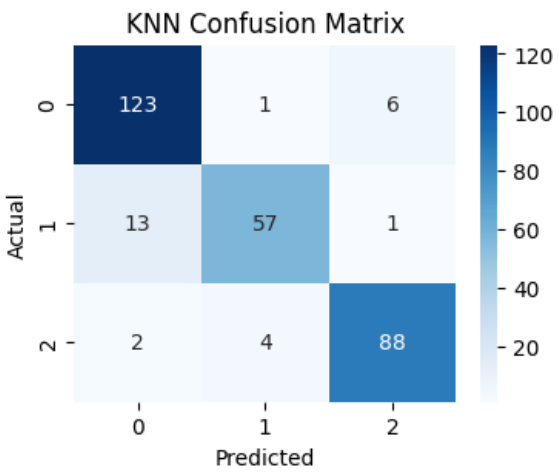
3.1.3. Median filter



Slika 9. Raspodela po kolonama nakon promene Median filtera

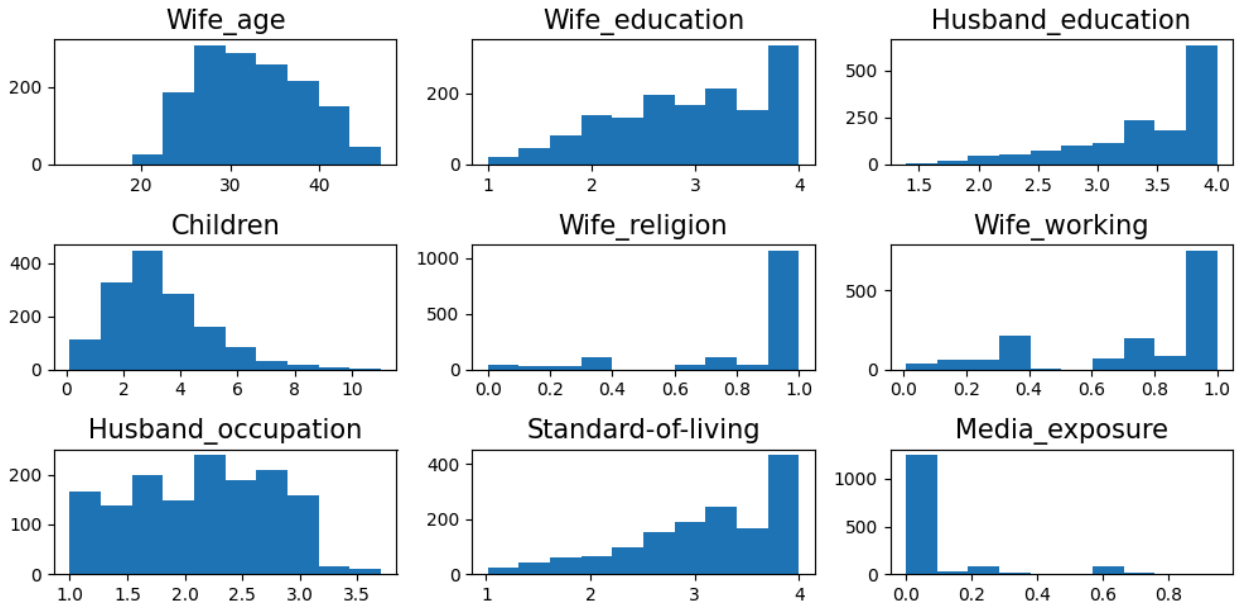


Slika 10. Random Forest nakon primene Median filtera

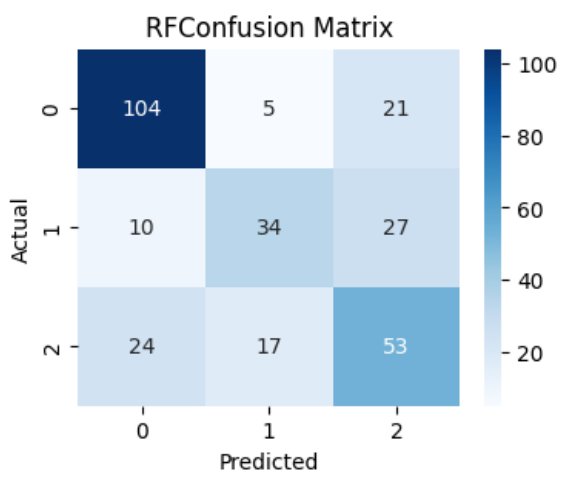


Slika 11. KNN nakon primene Median filtera

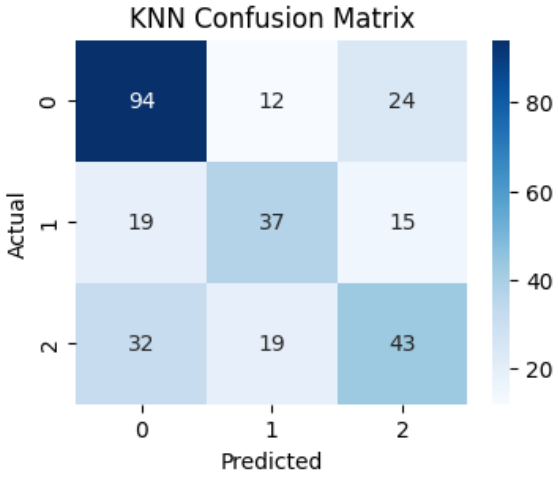
3.1.4. Kalaman filter



Slika 12. Raspodela po kolonama nakon primene Kalaman filtera

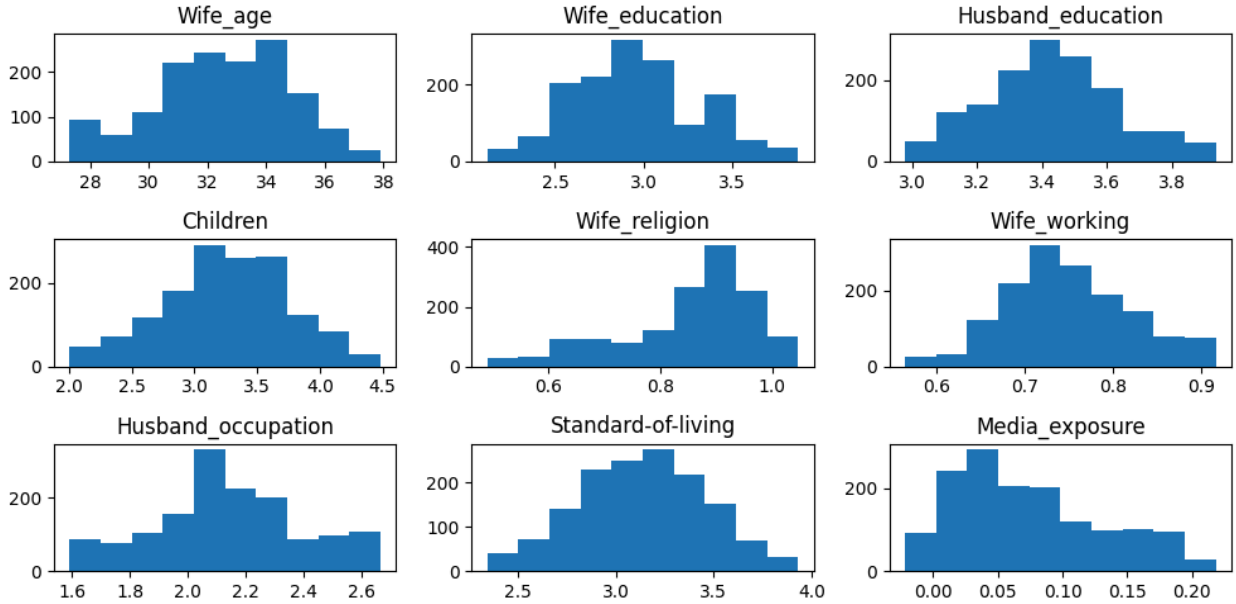


Slika 13. Random Forest nakon primene Kalaman filtera

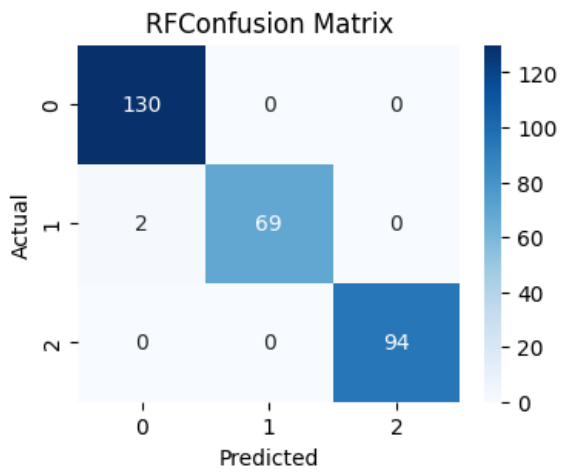


Slika 14. KNN nakon primene Kalaman filtera

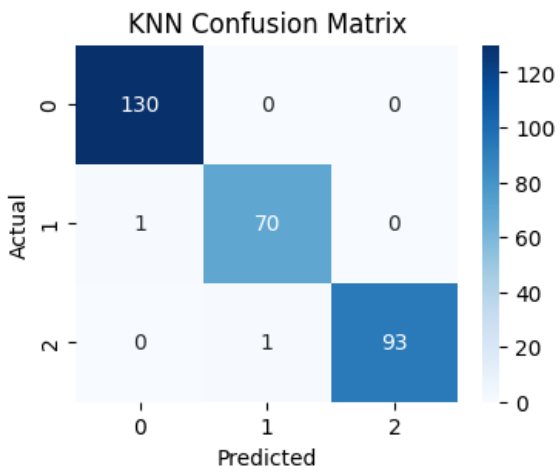
3.1.5. Fourijeva transformacija



Slika 15. Raspodela po kolonama nakon primene Fouriejove transofrmacije

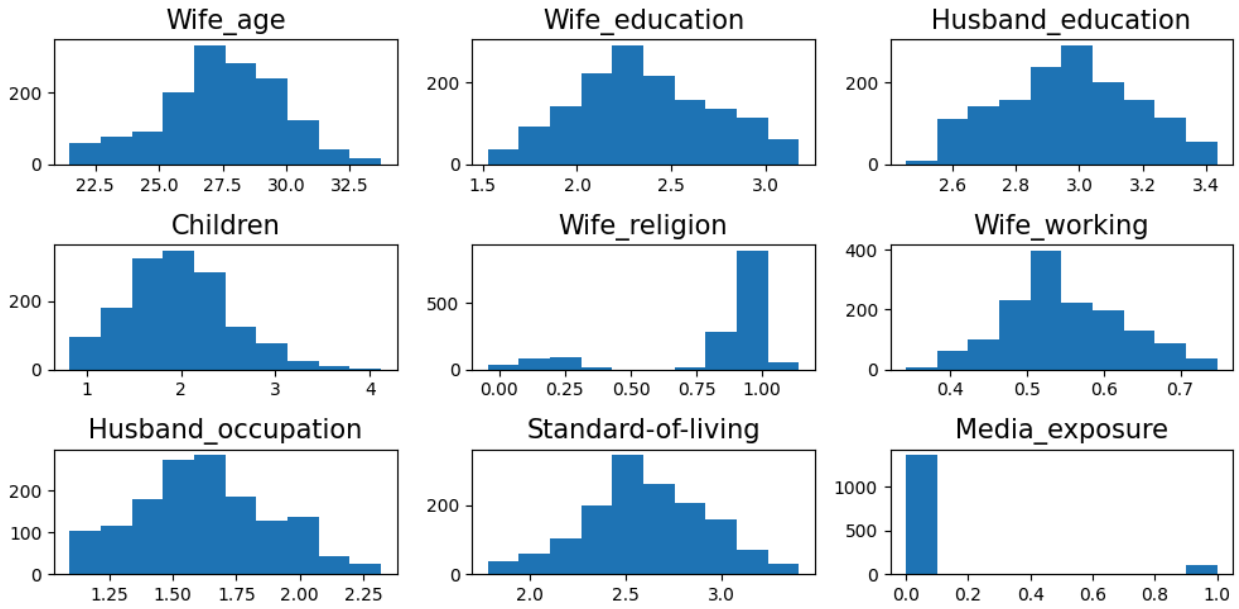


Slika 16. RF nakon primene Fourijeove transformacije

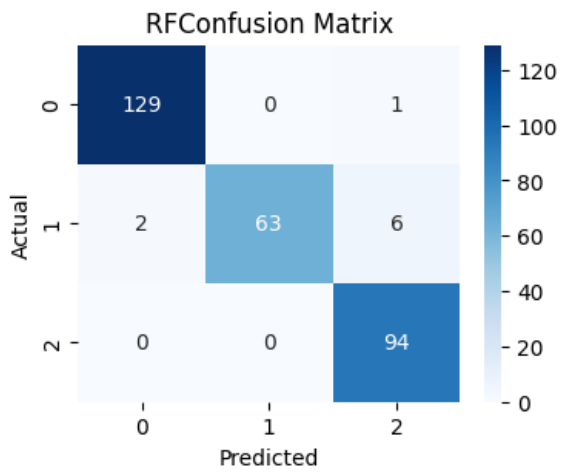


Slika 17. KNN nakon primene fourijeove transformacije

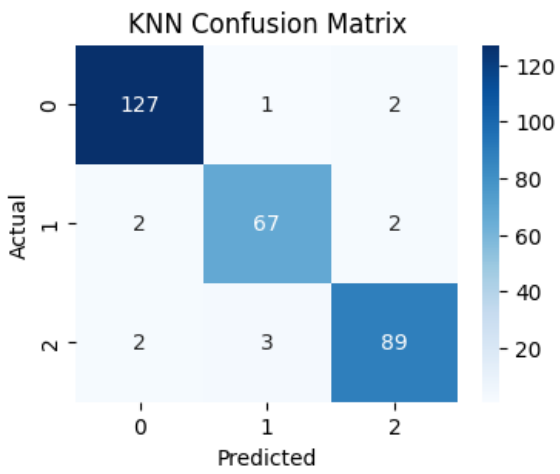
3.1.6. Welvet transformacija



Slika 18. Raspodela po kolonama nakon primene Welvet tranformacije

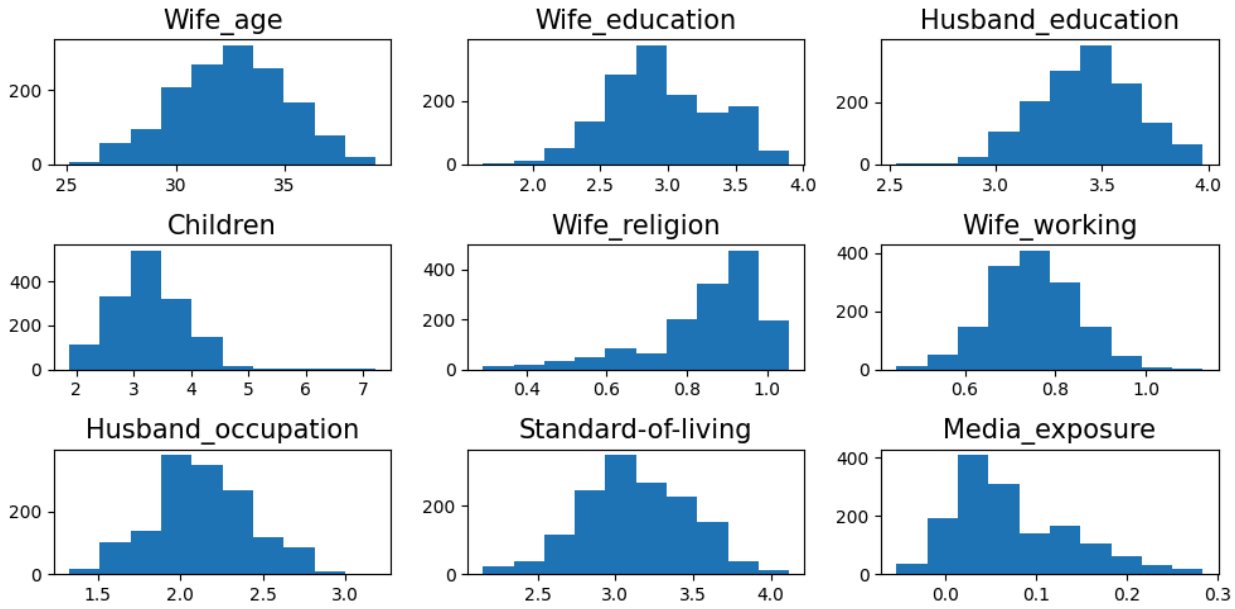


Sika 19. Random Forest nakon primene Welvet transformacije

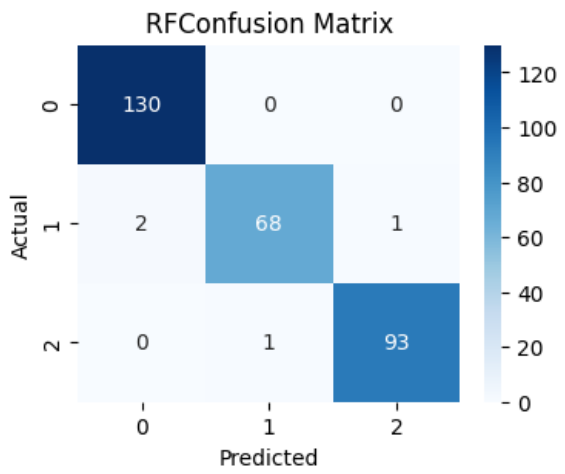


Slika 20. KNN nakon primene Welvet transformacije

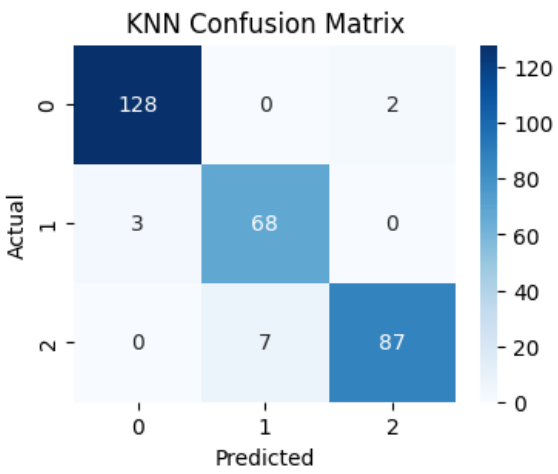
3.1.7. Savitzky-Golay filter



Slika 21. Raspodela po koonama nakon primene Savitzky-Golaz filtera



Slika 22. Random Forest nakon primene Savitzky-Golaz filtera



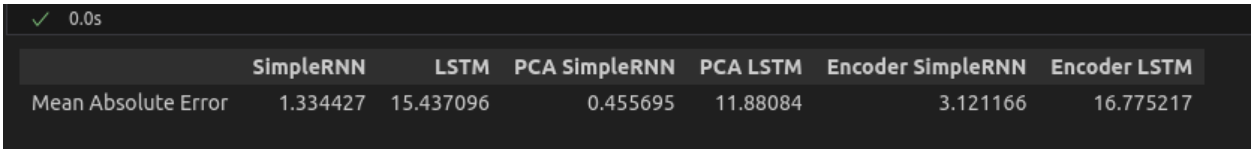
Slika 23. KNN nakon primene Savitzky-Golaz filter

3.2. Metode mašinskog učenja

Za demonstraciju metoda mašinskog učenja korišćen je wind energy dataset sa Kaggle platforme, koji sadrži informacije o proizvonji električne energije prikupljene od 4 nemačkih kompanija koje se bave proizvodnjom struje. Sadrži podatke o proizvodnji električne energije (nenormalizovane) sa intervalom od 15 minuta, ukupno 96 zapisa dnevno.

Korišćena su dva modela LSTM i Simple RNN koji su primenjeni prvo nad originalnim podacima sa šumom, a zatim i nad podacima kojima je redukovan šum primenom neke od prethodno opisanih metoda za redukciju.

Krajnji rezultat je prikazan na sledecoj slici.



	SimpleRNN	LSTM	PCA SimpleRNN	PCA LSTM	Encoder SimpleRNN	Encoder LSTM
Mean Absolute Error	1.334427	15.437096	0.455695	11.88084	3.121166	16.775217

Slika 24. Rezultat nakon metoda mašinskog učenja

3.2.1. GAN

Za demonstraciju korišćenja GAN mreže u redukciji šuma korišćen je MINST10 dataset koji sadrži fotografije brojeva pisanih rukom. Tim fotografijama je dodat šum. Na sledećoj slici je prikazan dataset pre i nakon dodavanja šuma.

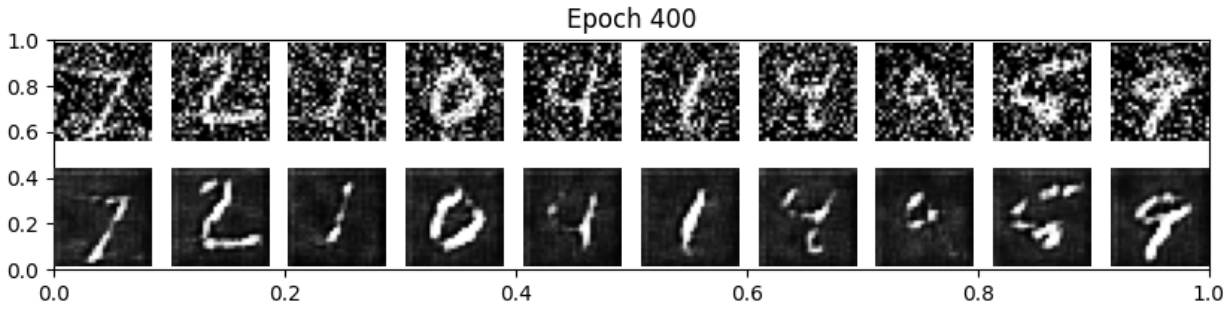
Generator se sastoji od sledećih slojeva:

1. Ulazni sloj (Input): Definiše ulazne dimenzije slike, ovde 28×28 sa jednim kanalom (siva skala).
2. Konvolucioni slojevi (Conv2D): Koriste filtere veličine 3×3 za ekstrakciju karakteristika iz slike. padding="same" omogućava očuvanje dimenzija slike, a relu aktivacija uvodi nelinearnost.
3. Batch Normalization: Normalizuje izlaze konvolucionih slojeva radi stabilnijeg i bržeg učenja, smanjujući verovatnoću prenaučavanja.
4. UpSampling2D: Povećava dimenzije slike bez promene rezolucije (ovde je faktor uvećanja 1×1, tako da dimenzije ostaju iste).
5. Izlazni sloj (Conv2D sa sigmoid aktivacijom): Proizvodi izlaz sa istim dimenzijama kao ulaz, dok sigmoid aktivacija ograničava vrednosti piksela između 0 i 1, korisno za slike u sivim tonovima.

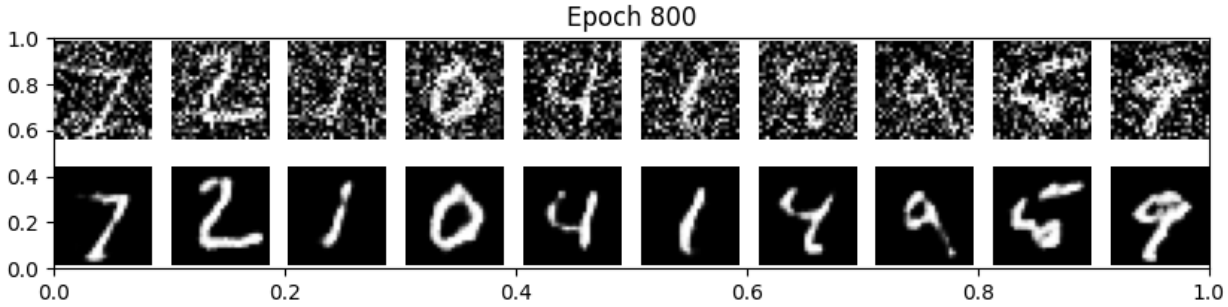
Deskriminator se sastoji od sledećih slojeva:

1. Ulazni sloj (Input): Definiše oblik ulazne slike kao 28××28 sa jednim kanalom (crno-bela slika).
2. Prvi konvolucioni sloj (Conv2D): Koristi 64 filtera veličine 3×3 sa korakom 2×2, što smanjuje dimenzije slike na polovinu (sa 28×28 na 14×14). padding="same" zadržava iste dimenzije slike na izlazu, a relu aktivacija uvodi nelinearnost.
3. Batch Normalization: Normalizuje izlaze prethodnog konvolucionog sloja kako bi stabilizovao učenje i ubrzao konvergenciju.
4. Drugi konvolucioni sloj (Conv2D): Ima 128 filtera veličine 3×3 sa korakom 2×2, što dodatno smanjuje dimenzije slike (sa 14×14 na 7×7). Aktivacija relu i padding="same" se ponovo koriste kako bi se sačuvale prostorne dimenzije.
5. Flatten: Transformiše višedimenzionalni izlaz prethodnog sloja u jedinstveni vektor, pripremajući podatke za potpuno povezani sloj.
6. Izlazni sloj (Dense sa sigmoid aktivacijom): Ima jedan neuron sa sigmoid aktivacijom koji vraća vrednost između 0 i 1, što odgovara verovatnoći jedne od dve klase u binarnoj klasifikaciji.

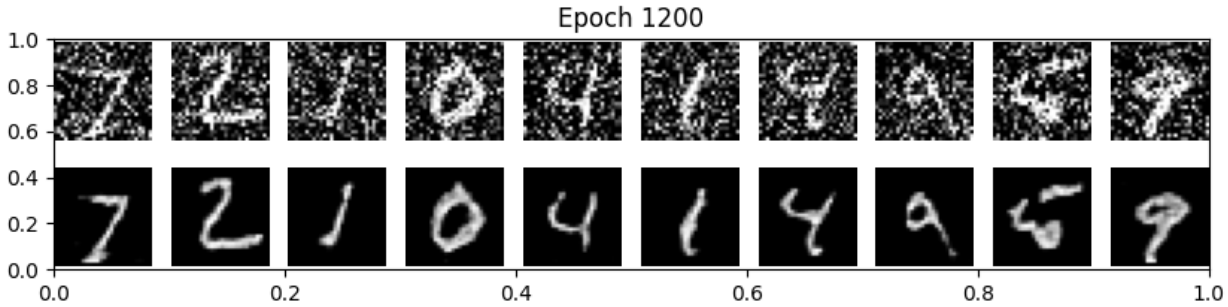
Model je treniran sa 2000 epoha i 64 batch size, s tim što se na svakih 400 epoha istrenirani model čuvao u fajl. Dakle postoji 5 modela nakon 400, 800, 1200, 1600, 2000 epoha.



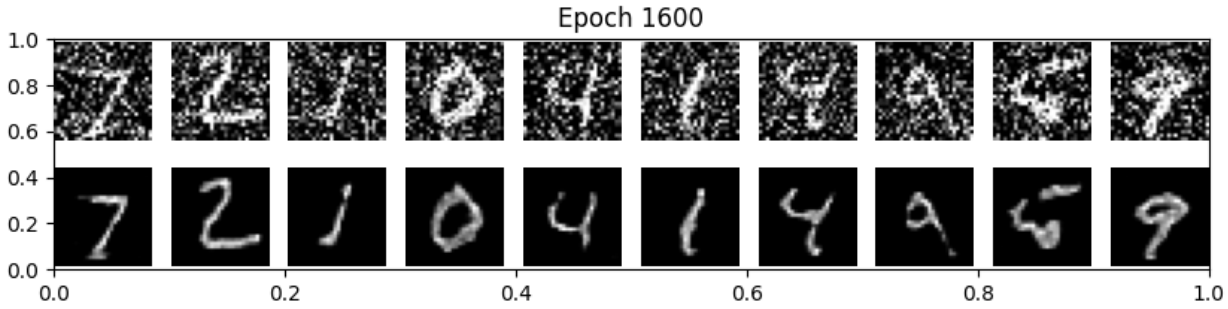
Slika 25. GAN sa 400 epoha



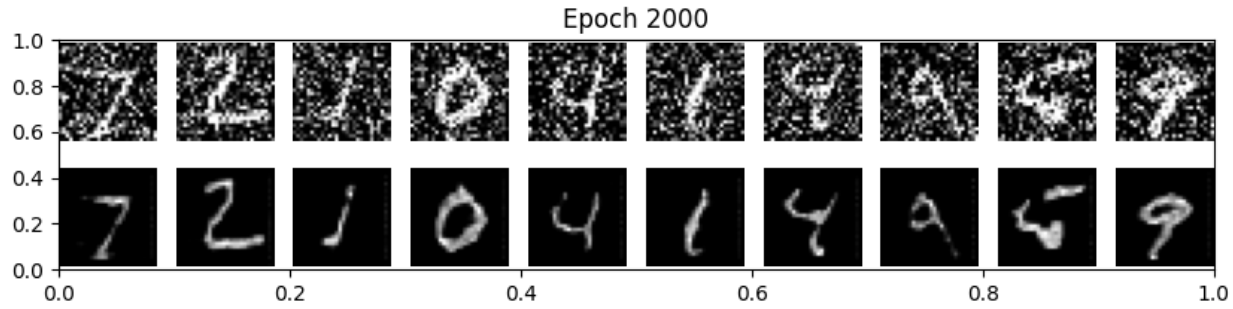
Slika 26. GAN sa 800 epoha



Slika 27. GAN sa 1200 epoha



Slika 28. GAN sa 1600 epoha



Slika 29. GAN sa 2000 epoha

Iz primera se jasno može videti da se najčistiji rezultati dobijaju sa modelom treniranim na 800 epoha. U slučaju sa 2000 epoha dolazi do pretreniranosti modela koji ne raspoznaje baš najbolje brojeve i i sam model počinje da generiše šum.

Zaključak

Šum u podacima predstavlja značajan izazov u analizi i obradi informacija, jer može negativno uticati na tačnost i pouzdanost klasifikacionih modela. Kroz ovaj rad, istraženi su različiti tipovi šuma, njihovi izvori i posledice na kvalitet podataka, kao i razlika između šuma i anomalija. Posebna pažnja posvećena je analizi uticaja šuma na performanse klasifikacionih sistema, što je od ključne važnosti za primenu modela u realnim situacijama.

Prikazani metodi za uklanjanje šuma obuhvataju tradicionalne statističke pristupe, metode obrade signala i tehnike zasnovane na mašinskom učenju. Tradicionalne metode, poput medijan filtra i Kalmanovog filtra, pokazale su se efikasnim za osnovne probleme sa šumom, dok su naprednije metode, poput Fourierove i wavelet transformacije, omogućile preciznije uklanjanje kompleksnijih šumova. Sa druge strane, tehnike mašinskog učenja, poput PCA, autoenkodera i GAN modela, pružaju fleksibilnost i moćnu mogućnost za prepoznavanje i uklanjanje šuma u složenim strukturama podataka.

Praktična primena ovih metoda potvrđuje njihovu efikasnost, ali i ukazuje na značaj prilagođavanja tehnike specifičnostima skupa podataka. Iako nijedna metoda nije univerzalno primenjiva, kombinacija različitih pristupa može značajno unaprediti kvalitet podataka i performanse klasifikacionih modela.

Literatura

[1] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, Intelligent Systems Reference Library, Springer, 2015.

[2] V. Kumar, P.-N. Tan, M. Steinback, and A. Karpatne, *Introduction to Data Mining*, Pearson, 2019.

[3] "Noisy Data," *TechTarget*. Available: https://www.techtarget.com/searchbusinessanalytics/definition/noisy-data#:~:text=Noisy%20data%20is%20a%20data,a%20higher%20quality%20data%20set.

[4] "Noisy Data," *SCI2S UGR*. Available: https://sci2s.ugr.es/noisydata.

[5] R. Y. Wang, V. C. Storey, and C. P. Firth, "A framework for analysis of data quality research," *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, no. 4, pp. 623-640, 1995. doi: 10.1109/69.404034.

[6] X. Zhu and X. Wu, "Class Noise vs. Attribute Noise: A Quantitative Study," *Artificial Intelligence Review*, vol. 22, pp. 177-210, 2004. doi: 10.1007/s10462-004-0751-8.

[7] "Noisy Data," *SCI2S UGR*. Available: https://sci2s.ugr.es/noisydata.

[8] "How to Handle Noise in Machine Learning," *GeeksforGeeks*. Available: https://www.geeksforgeeks.org/how-to-handle-noise-in-machine-learning/.

[9] "What's That Noise? Of Systematic and Random Errors or Why Pilot Testing Is Important," *Prolific Researcher Help*. Available: https://researcher-help.prolific.com/hc/en-gb/articles/360009377894--What-s-that-noise-Of-systematic-and-random-errors-or-Why-pilot-testing-is-important.

[10] "What Is the Difference Between Noise and Outliers?" *Quora*. Available: https://www.quora.com/What-is-the-difference-between-noise-and-outliers.

[11] "The Basic Difference Between Noise and Outliers in Data," *Medium*. Available: https://medium.com/@aatl2012/the-basic-difference-between-noise-and-outliers-in-data-cd3ff32343e0.

[12] "Handling Noisy Data: Smoothing and Filtering Techniques," *Data Headhunters*. Available: https://dataheadhunters.com/academy/handling-noisy-data-smoothing-and-filtering-techniques/.

[13] "How to Handle Noise in Machine Learning," *GeeksforGeeks*. Available: https://www.geeksforgeeks.org/how-to-handle-noise-in-machine-learning/.

[14] "How to Use Machine Learning to Separate the Signal from the Noise," *Skan*. Available: https://www.skan.ai/blogs/how-to-use-machine-learning-to-separate-the-signal-from-the-noise-skan.

[15] "Guide to Contrastive Learning," *Encord*. Available: https://encord.com/blog/guide-to-contrastive-learning/.