UNIVERSITATEA „ALEXANDRU IOAN CUZA” DIN IAȘI FACULTATEA DE ECONOMIE ȘI ADMINISTRAREA AFACERILOR
SPECIALIZAREA: DATA MINING

Clasificarea salariului in funcție de caracteristicile socio-demografice a individului cu ajutorul arborilor de decizie

COORDONATOR ȘTIINȚIFIC:
PROF. DR. CIPRIAN IONEL TURTUREAN

STUDENT:
COZMA ALEXANDRU-CRISTIAN

IAȘI
2024

Cuprins

Introducere

Analiza salariilor în funcție de caracteristicile socio-demografice ale indivizilor constituie un subiect de interes major în domeniile economiei și științelor sociale, oferind o perspectivă esențială asupra modului în care factori precum educația, experiența, genul sau sectorul economic influențează nivelul veniturilor și generează inegalități pe piața muncii. În acest context, studiul de față își propune să investigheze aceste relații în cadrul unui cadru istoric specific – Statele Unite ale Americii în anul 1976 – utilizând un set de date extras din chestionarul *Current Population Survey*, care conține informații detaliate despre 526 de respondenți.

Obiectivul principal al cercetării este de a construi și evalua modele predictive capabile să clasifice salariul orar al indivizilor în funcție de o serie de variabile explicative, raportându-se la un prag de referință stabilit la valoarea medie a salariului orar din anul 1976, respectiv 5,58 dolari. Pentru realizarea acestui demers sunt utilizate tehnici moderne de modelare predictivă bazate pe arbori de decizie, prin aplicarea a trei algoritmi consacrați: **CART, QUEST și CHAID**. Alegerea acestor metode este justificată de avantajele lor complementare: CART oferă o structură binară clară și interpretabilă, QUEST optimizează selecția variabilelor reducând bias-ul, iar CHAID evidențiază relațiile statistice semnificative dintre variabile prin utilizarea testelor de asociere.

Pentru a asigura calitatea datelor și robustețea modelelor construite, analiza presupune o etapă riguroasă de preprocesare, care include eliminarea variabilelor irelevante, discretizarea variabilelor continue, transformarea variabilelor dummy în variabile categoriale și tratarea valorilor extreme identificate în urma analizei exploratorii. Aceste transformări permit o modelare mai eficientă și reduc riscul de supraîncărcare sau distorsionare a algoritmilor.

Prin compararea performanțelor celor trei algoritmi aplicați, studiul urmărește să evidențieze care dintre metode oferă cele mai bune rezultate predictive și care sunt variabilele cu cel mai mare impact asupra salariului. În plus, rezultatele obținute pot aduce o contribuție valoroasă la înțelegerea mecanismelor care stau la baza inegalităților salariale, oferind totodată un suport empiric pentru fundamentarea unor politici economice și sociale mai echitabile. Deși analiza se desfășoară într-un cadru temporal limitat, concluziile sale pot avea relevanță și în contexte contemporane similare.

Prin aplicarea unor metode avansate de învățare automată asupra unui set de date clasic, dar semnificativ din punct de vedere socio-economic, această lucrare își propune să contribuie la aprofundarea cunoașterii în domeniul analizei veniturilor și să ofere o bază solidă pentru cercetări viitoare în sfera economiei muncii și a politicilor public.

1. Prezentarea bazei de date

Baza de date care sta la baza acestui proiect este reprezentata de răspunsurile înregistrate in cadrul chestionarului **Current Population Survey** din 1976. Acest studiu ilustrează diferite caracteristici socio-demografice ale unui eșantion de 526 de cetățeni din SUA.

1.1. Descrierea bazei de date inițiale

Pentru fiecare dintre cei 526 de respondenți au fost măsurate 25 de caracteristici, acestea sunt:

- **rownames:** variabila de indexare
- **wage:** salariul pe ora **(variabila obiectiv)**
- **educ:** anii de educație
- **exper:** anii de experiență profesionala
- **tenure:** anii petrecuți cu angajatorul curent cu contract pe perioadă nedeterminată
- **nonwhite:** = minoritate rasiala (1-minoritate, 0- alb)
- **female:** = gen (1-femeie, 0-barbat)
- **married:** = statut civil (1-casatorit, 0-necasatorit)
- **numdep:** numărul de dependenți
- **smsa:** = mediul de proveniența (1-urban, 0-rural)
- **northcen:** = traieste in nord-centrul SUA
- **south:** = traieste in sudul SUA
- **west:** =1 traieste in vestul SUA
- **construc:** = lucreaza in constructii
- **ndurman:** = lucreaza in industria de producție a bunurilor ne-durabile
- **trcommpu:** = lucreaza in transport public
- **trade:** = lucreaza in industria de comert cu amanuntul
- **services:** = lucreaza in industria serviciilor
- **profserv:** = lucreaza in industria serviciilor profesionale.
- **profocc:** = are o ocupatie profesionala
- **clerocc:** = este sau nu cleric
- **servocc:** = ocupatie in domeniul serviciilor
- **lwage:** log(wage)
- **expersq:** exper^2
- **tenursq:** tenure^2

Output 1.1: Structura setului de date

```
> glimpse(wage1)
Rows: 526
Columns: 25
$ rownames <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 2…
$ wage     <dbl> 3.10, 3.24, 3.00, 6.00, 5.30, 8.75, 11.25, 5.00, 3.60, 18.18, 6.25, 8.13, 8.77, 5.50, 22.20, 17.33, 7.50…
$ educ     <int> 11, 12, 11, 8, 12, 16, 18, 12, 12, 17, 16, 13, 12, 12, 12, 16, 12, 13, 12, 12, 12, 12, 16, 12, 11, 16, 1…
$ exper    <int> 2, 22, 2, 44, 7, 9, 15, 5, 26, 22, 8, 3, 15, 18, 31, 14, 10, 16, 13, 36, 11, 29, 9, 3, 37, 3, 11, 31, 30…
$ tenure   <int> 0, 2, 0, 28, 2, 8, 7, 3, 4, 21, 2, 0, 0, 3, 15, 0, 0, 10, 0, 6, 4, 13, 9, 1, 8, 3, 10, 0, 0, 1, 5, 5, 16…
$ nonwhite <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ female   <int> 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1,…
$ married  <int> 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0,…
$ numdep   <int> 2, 3, 2, 0, 1, 0, 0, 0, 2, 0, 0, 0, 2, 0, 1, 1, 0, 0, 3, 0, 0, 3, 0, 0, 0, 1, 1, 0, 2, 2, 2, 2, 0, 0, 0,…
$ smsa     <int> 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ northcen <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ south    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ west     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ construc <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ ndurman  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ trcommpu <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,…
$ trade    <int> 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,…
$ services <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1,…
$ profserv <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,…
$ profocc  <int> 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0,…
$ clerocc  <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,…
$ servocc  <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ lwage    <dbl> 1.1314021, 1.1755733, 1.0986123, 1.7917595, 1.6677068, 2.1690538, 2.4203682, 1.6094379, 1.2809339, 2.900…
$ expersq  <int> 4, 484, 4, 1936, 49, 81, 225, 25, 676, 484, 64, 9, 225, 324, 961, 196, 100, 256, 169, 1296, 121, 841, 81…
$ tenursq  <int> 0, 4, 0, 784, 4, 64, 49, 9, 16, 441, 4, 0, 0, 9, 225, 0, 0, 100, 0, 36, 16, 169, 81, 1, 64, 9, 100, 0, 0…
```

Conform Outputului 1.1, setul „wage1” conține 5 variabile numerice utile, care pot fi folosite in procesul de modelare și anume salariul pe ora (wage), anii de educație (educ), anii de experiența pe piața muncii (exper) și perioada de angajare pe contract nedeterminat (tenure). Mai există 12 variabile dummy care pot fi transformate ulterior in 2 variabile categoriale care indica tipul de job pe care il practica individul si regiunea în care rezidează. Pe lângă acestea mai sunt prezente 4 variabile categoriale utile și anume rasa (nonwhite), female (gen), statutul marital (married), numărul de persoane dependente (numdep) și mediu de proveniență (smsa). Restul variabilelor rămase țin de modelele econometrice realizate de cel care a colectat datele (lwage, expersq, tenures) plus o variabilă de indexare (rownumber).

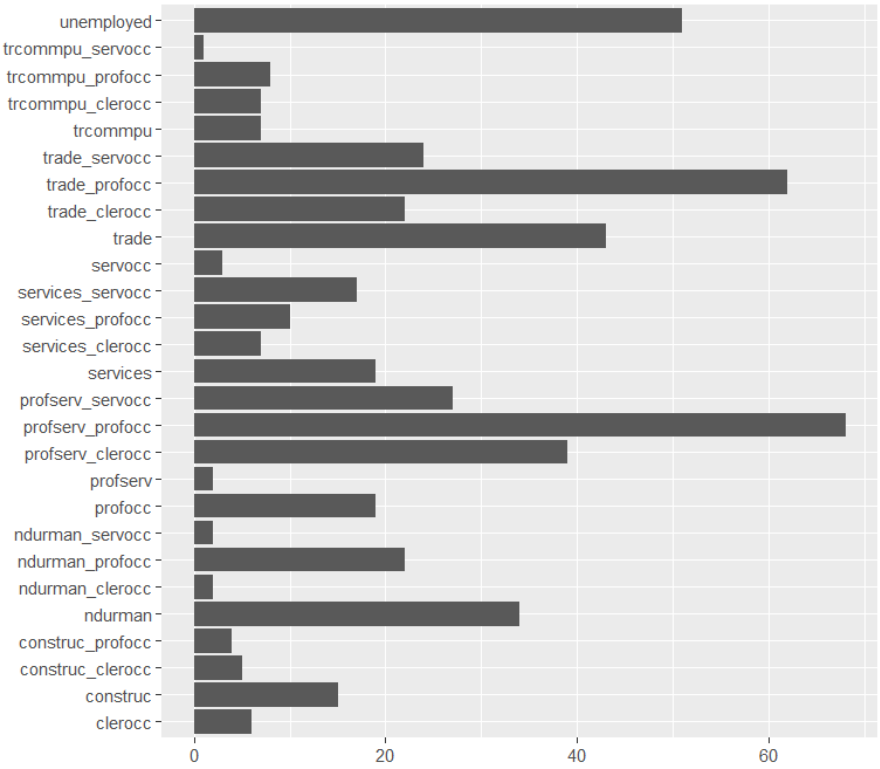
1.2. Operațiuni preliminare

Întrucât setul inițial prezintă deficiențe de structură prin prisma existenței variabilelor dummy și a variabilelor inutile, acest subcapitol are ca obiectiv explicarea metodelor de ameliorare a acestor deficiențe prin eliminarea variabilelor inutile și definirea a 2 noi variabile categoriale menite să ilustreze informațiile reliefate în cele 12 variabile dummy.

9 din cele 12 variabile dummy reflectă profesia și/sau ocupația unui individ, cei cu valoarea 0 pentru toate categoriile de activități economice, au fost încadrate în clasa „unemployed”. În Graficul 1.1 este reprezentată combinația dintre industria economică și tipul ocupației exercitate. Cele mai frecvent întâlnite categorii sunt „profserv_profocc” și „profserv_clerocc”, indicând o concentrare a forței de muncă în serviciile profesionale, atât în poziții de specialitate, cât și clericale. De asemenea, se observă o proporție ridicată a persoanelor „unemployed”, ceea ce

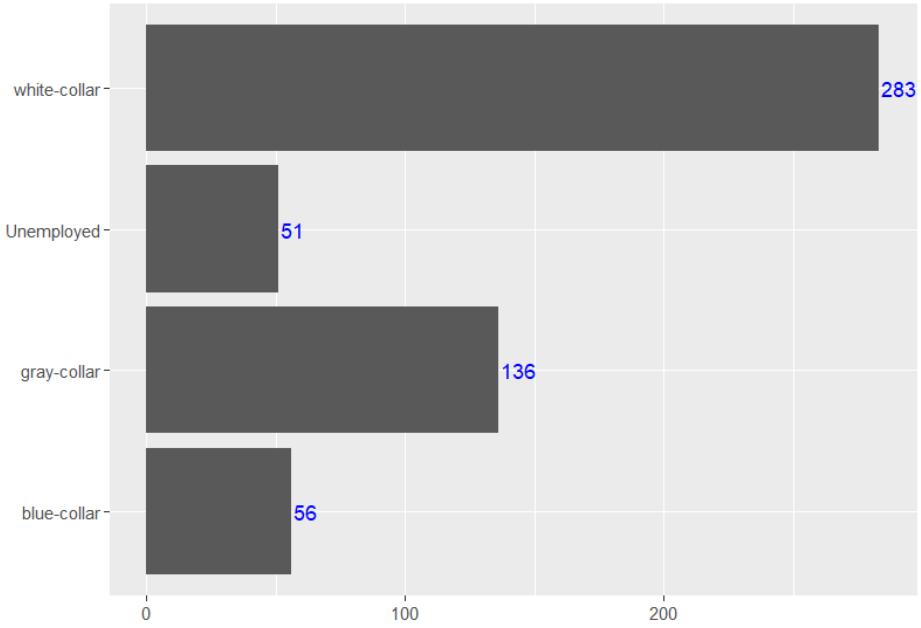
evidențiază un posibil dezechilibru pe piața muncii. Sectoare precum comerțul cu amănuntul („trade”) și industria bunurilor ne-durabile („ndurman”) sunt moderat reprezentate, în timp ce industrii precum construcțiile, transportul public sau serviciile generale au ponderi reduse. Distribuția sugerează o polarizare ocupațională, fiind dominată de profesii specializate și de un segment relevant fără activitate economică, aspecte utile în etapa de modelare și selecție a variabilelor pentru analiza predictivă.

Graficul 1.1: Distribuția indivizilor în funcție de specificul activității economice practicate



Cu un total de 28 de categorii, variabila nou creată „econsect” are o structură complexă care ar pute fi simplificată. Küsbeci & Urgan clasifică în 2022 toate posibilele tipologii de activități în funcție de natura muncii. Principalele categorii identificate, care pot fi create luând în calcul caracterul vag al informațiilor din setul de date sunt „blue collar”, „white collar” și „gray collar” metodă pe care am folosit-o pentru a simplifica variabila „econsect”. Graficul 1.2 evidențiază o predominanță clară a ocupațiilor white-collar (283 indivizi), urmate de gray-collar (136), ceea ce sugerează o forță de muncă orientată spre activități intelectuale sau mixte. Ocupațiile blue-collar (56) și persoanele unemployed (51) sunt mai puțin reprezentate, indicând o pondere redusă a muncii manuale și un nivel relativ scăzut al șomajului în eșantion.

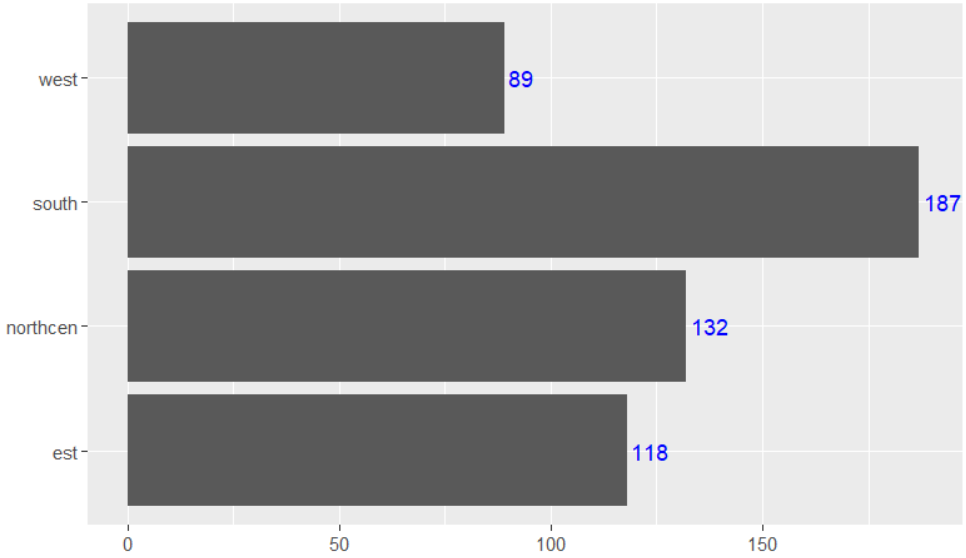
Graficul 1.2: Distribuția indivizilor în funcție de natura muncii



Meseriile care intra in categoria **blue-collar jobs (BC)** sunt: **construc** (Construcții), **ndurman** (Industria manufacturieră de bunuri de consum), **trcommpu** (Transport, telecomunicații, utilități publice). Meseriile care intra in categoria **white-collar jobs (WC)** sunt: **profserv** (Servicii profesionale), **profocc** (Ocupații profesionale), **clerocc** (Ocupații clericale), **profserv_profocc, profserv_clerocc, ndurman_profocc, trcommpu_profocc, trade_profocc, services_profocc, services_clerocc, trade_clerocc, construc_profocc, construc_cleroc**c. Meseriile care intra in categoria **gray-collar jobs (GC)** sunt: **trade** (Comerț en-gros și en-detail), **services** (Industria serviciilor), **servocc** (Ocupații din domeniul serviciilor), **trade_servocc, services_servocc, construc_servocc, ndurman_servocc, trcommpu_servocc.**

A doua variabilă creată din variabile dummy este variabila „region”. Acesta este compusă din „east”, „south” și „northcen”, indivizii care au înregistrat valoarea 0 pentru toate cele 3 variabile au fost categorisiți ca fiind din vestul țării. Graficul 1.3 ilustrează distribuția variabilei „region”. Regiunea south este cea mai reprezentată (187 indivizi), urmată de northcen (132) și est (118), în timp ce west are cea mai mică pondere (89). Această distribuție poate influența analiza, sugerând posibile diferențe regionale în ceea ce privește ocuparea, veniturile sau accesul la resurse economice.

Grafic 1.3: Distribuția indivizilor în funcție de regiunea geografică



2. Metodologie

Atât pentru partea de procesare a datelor cât și pentru partea de modelare am folosit **R**. Pachetul de baza a fost folosită pentru operațiunile preliminare, **ggplot2** a fost folosit pentru realizarea graficelor, pentru discretizare am folosit pachetul **discretisation** și **corplot** si **rcompanion** pentru alegerea variabilelor. În ceea ce privește partea de modelare, am ales **algoritmii CART, QUEST și CHAID** care au fost construite cu ajutorul pachetelor **rpart, caret, partykit și CHAID.** Acești algoritmi sunt recunoscuți pentru abilitatea lor de a clasifica eficient datele, fiecare având avantaje specifice în analiza complexă a salariului.

3. Analiza exploratorie a datelor

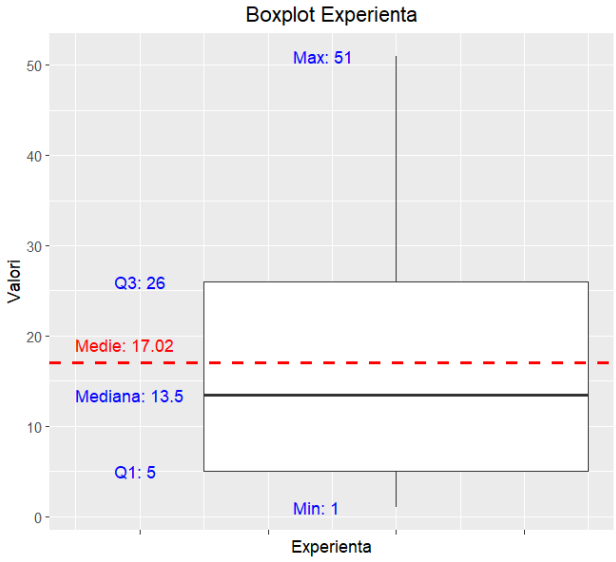
Analiza exploratorie a datelor reprezintă o etapă esențială în procesul de prelucrare și modelare a datelor, având ca scop înțelegerea structurii interne a setului de date, identificarea tiparelor, relațiilor și eventualelor anomalii. Prin intermediul statisticii descriptive și al vizualizărilor grafice, această secțiune urmărește evidențierea caracteristicilor variabilelor, evaluarea distribuțiilor și detectarea valorilor lipsă sau extreme, pregătind astfel datele pentru etapele ulterioare de procesare și modelare predictivă.

3.1. Statistica descriptivă și vizualizarea datelor

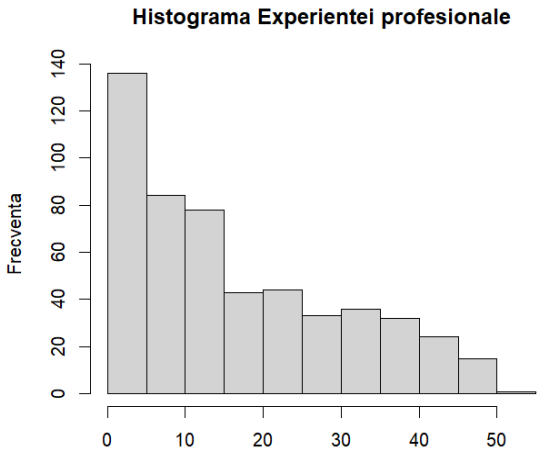
3.1.1 Variabile numerice

Acest subcapitol se va concentra pe cele 5 variabile specifice din setul de date inclusiv variabila obiectiv, aceste variabile vor fi ulterior discretizate pentru a se plia pe modul de funcționare a algoritmului CHAID, care poate fi construit doar cu variabile discrete.

Graficul 3.1



Graficul 3.2

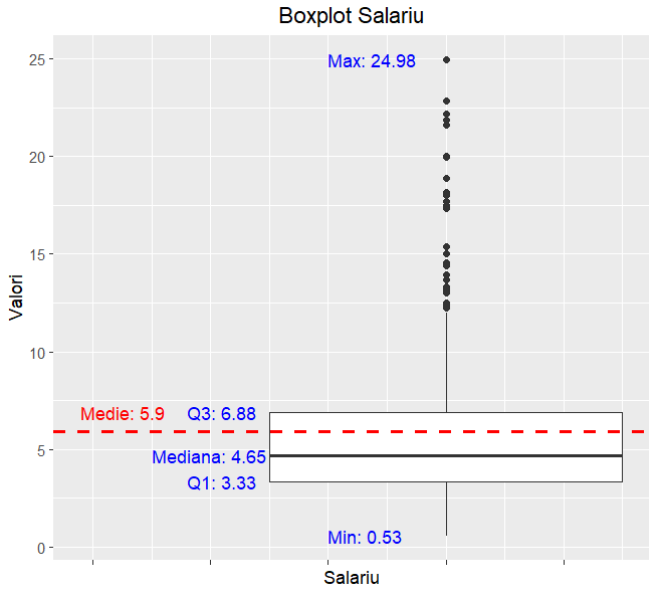


Pentru prima variabila, experiența profesională, graficele 3.1 și 3.2 ilustrează anumite informați, și nume:

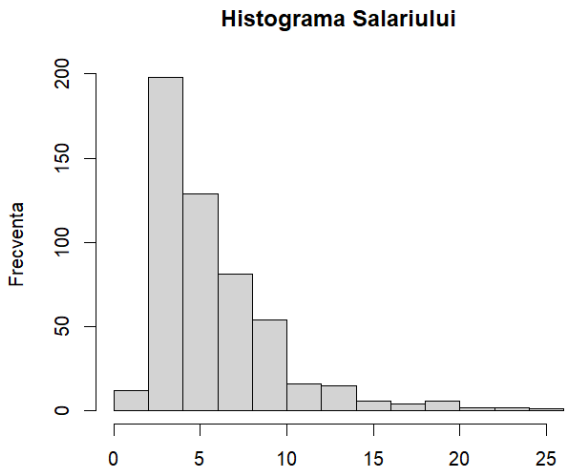
- **Min**: La nivelul esantionului, cel mai mic numar de ani petrecuti de catre un individ pe piata muncii este 1
- **Q1**: Primii 25% din indivizi au sub 5 ani de experiente, restul de 75% au mai mult de atat
- **Q2**: Primii 50% din indivizi au sub 13.5 ani de experiente, restul de 50% au mai mult de atat
- **Q3**: Primii 75% din indivizi au sub 26 ani de experiente, restul de 25% au mai mult de atat
- **Max**: : La nivelul esantionului, cel mai mare numar de ani petrecuti de catre un individ pe piata muncii este 26
- **Medie**: In medie, un individ din esantion are 17.02 ani de experienta

- **Normalitate**: Din histograma se poate observa ca distributia variabilei este asimetrica la dreapta si leptocurtica

Graficul 3.3



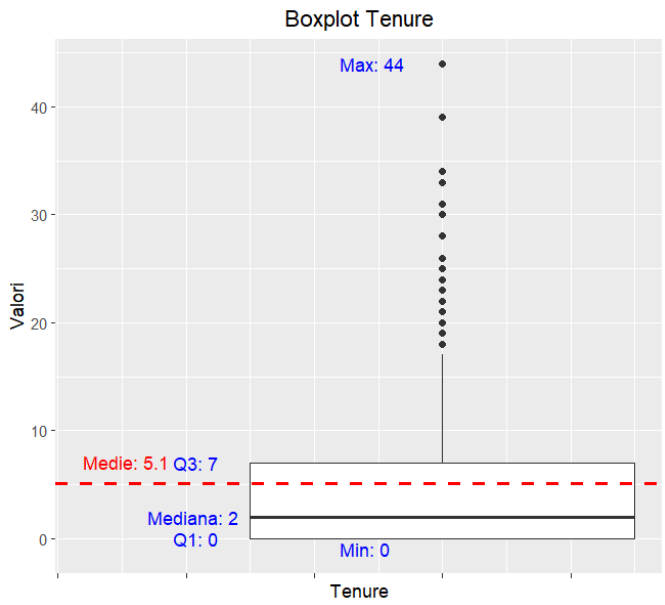
Graficul 3.4



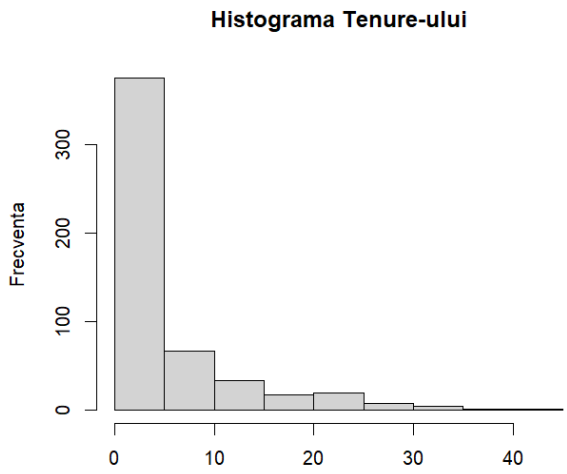
Salariul, din poziția de variabilă obiectiv, reprezintă singurul atribut care va fi discretizat cu certitudine. Din graficul 3.3 si 3.4 reiese:

- **Min**: La nivelul esantionului, cel mai mic salariu mediu pe ora a unui individ este de 0.53$
- **Q1**: Primii 25% din indivizi au sub 3.33 $/h, restul de 75% au mai mult de atat
- **Q2**: Primii 50% din indivizi au sub 4.65 $/h, restul de 50% au mai mult de atat
- **Q3**: Primii 75% din indivizi au sub 6.88 $/h, restul de 25% au mai mult de atat
- **Max**: : La nivelul esantionului, cel mai mare salariu mediu pe ora este de 24.98 $
- **Medie**: In medie, un individ din esantion castiga 5.9$/h
- **Normalitate**: Din histograma se poate observa ca distributia variabilei este asimetrica la dreapta si leptocurtica

Graficul 3.5



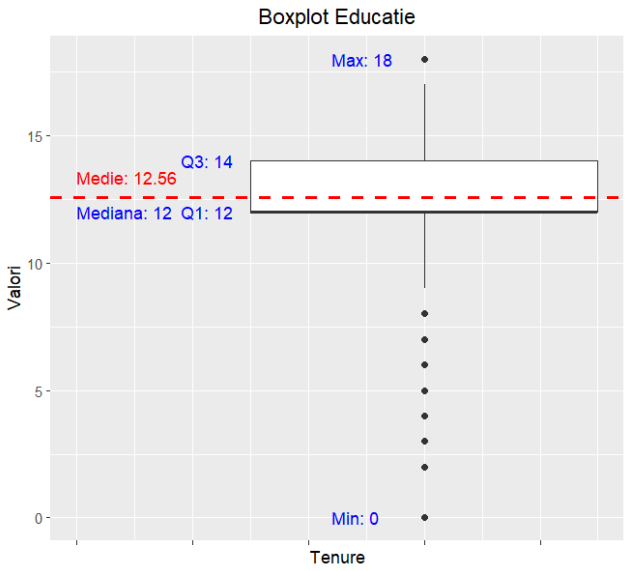
Graficul 3.6



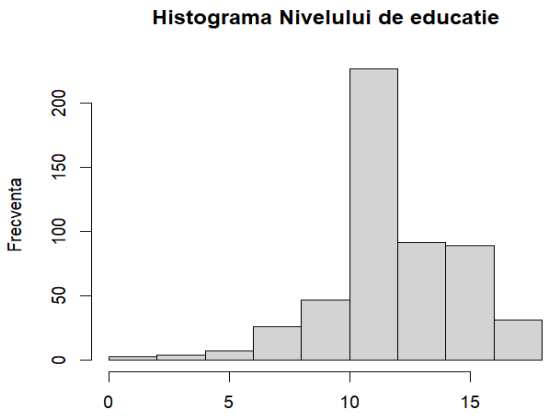
Și în cazul tenure, putem observa prezenta outlierilor deși mai puțini decât în distribuția salariului. Graficul 3.5 și 3.6 prezinta urmatoarele informatii:

- **Min**: La nivelul esantionului, individul cu cei mai putini ani consecutivi petrecuti cu angajatorul curent are valoarea 0.
- **Q1**: Primii 25% din indivizi au sub 0 ani petrecuti consecutivi cu acelasi angajator, restul de 75% au mai mult de atat
- **Q2**: Primii 50% din indivizi au sub 2 ani petrecuti consecutivi cu acelasi angajator, restul de 75% au mai mult de atat
- **Q3**: Primii 75% din indivizi au sub 7 ani petrecuti consecutivi cu acelasi angajator, restul de 75% au mai mult de atat
- **Max**: : La nivelul esantionului, individul cu cei mai multi ani consecutivi petrecuti cu angajatorul curent are valoarea 44.
- **Medie**: In medie, un individ din esantion a petrecut 5.1 ani consecutivi cu angajatorul curent.
- **Normalitate**: Din histograma se poate observa ca distributia variabilei este asimetrica la dreapta si leptocurtica

Graficul 3.7



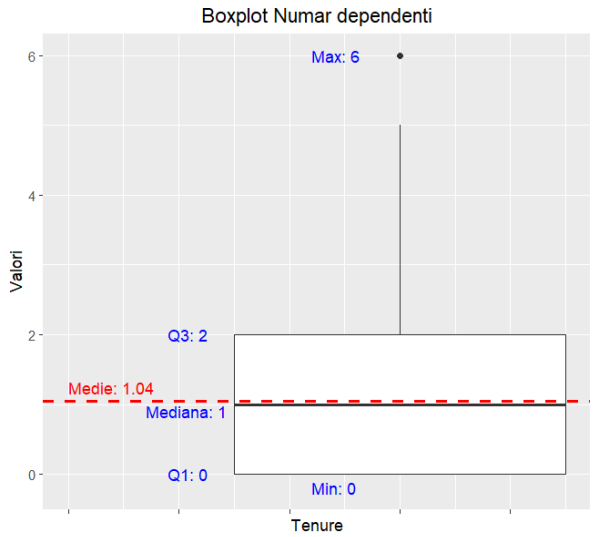
Graficul 3.8



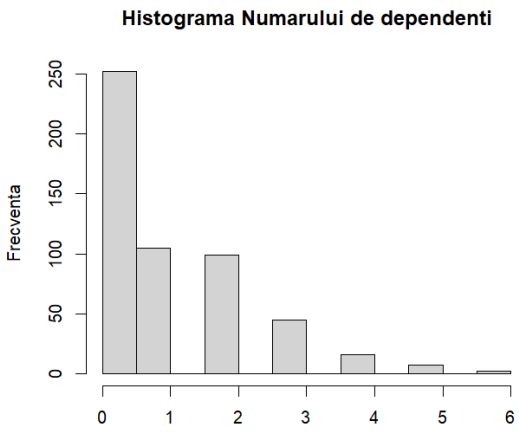
Și în cazul nivelului de educație sunt prezente valori extremă însă numărul acestora este neglijabil, alte informații despre distribuția valorilor sunt:

- **Min**: La nivelul esantionului, cel mai mic numar de ani de educatie este 0
- **Q1**: Primii 25% din indivizi au sub 12 ani de educatie, restul de 75% au mai mult de atat
- **Q2**: Primii 50% din indivizi au sub 12 ani de educatie, restul de 50% au mai mult de atat
- **Q3**: Primii 75% din indivizi au sub 14 ani de educatie, restul de 25% au mai mult de atat
- **Max**: : La nivelul esantionului, cel mai mare numar de ani de educatie este 18
- **Medie**: In medie, un individ din esantion are 12.56 ani de educatie
- **Normalitate**: Din histograma se poate observa ca distributia variabilei este asimetrica la stanga si leptocurtica

Graficul 3.9



Graficul 3.10



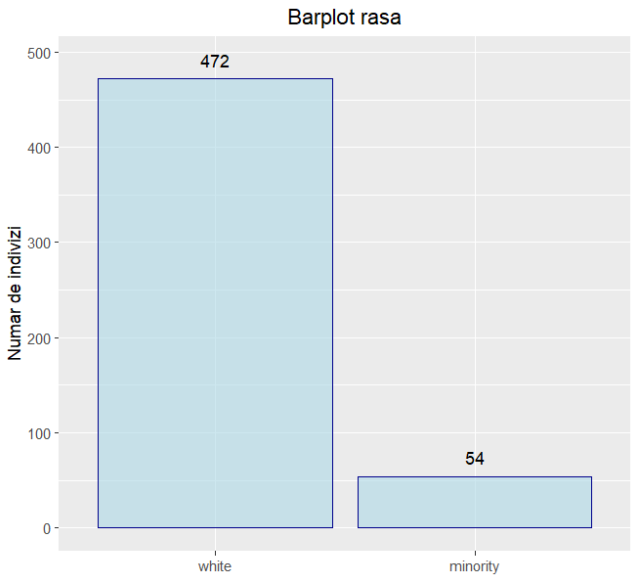
Numărul de dependenți nu prezintă decât o valoare extremă care este totuși realistă și nu trebuie neapărat înlăturată. În graficul 3.9 și 3.10 sunt ilustrate următoarele:

- **Min**: La nivelul esantionului, cel mai mic numar de persoane care depind de un respondent este 0
- **Q1**: Primii 25% din indivizi au 0 persoane care depind de acesta, restul de 75% au mai mult de atat
- **Q2**: Primii 50% din indivizi au 1 persoana care depinde de acesta, restul de 50% au mai mult de atat
- **Q3**: Primii 75% din indivizi au 2 persoane care depind de acesta, restul de 25% au mai mult de atat
- **Max**: : : La nivelul esantionului, cel mai mic numar de persoane care depind de un respondent este 6
- **Medie**: In medie, un individ din esantion are 1.04 persoane dependente
- **Normalitate**: Din histograma se poate observa ca distributia variabilei este asimetrica la stanga si leptocurtica

3.1.2 Variabile categoriale

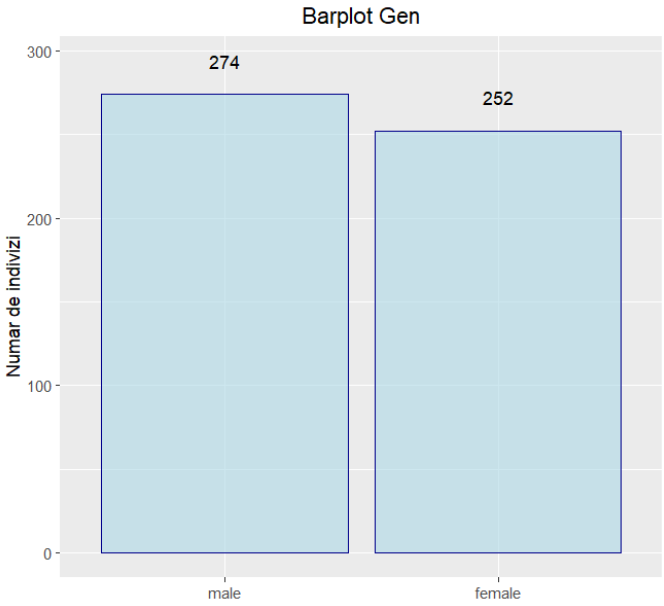
Pe lângă cele 2 variabile prezentate în primul capitol, am analizat distributiile altor 4 variabile din care se vor selecta doar cele relevante pentru modelare

Grafic 3.11



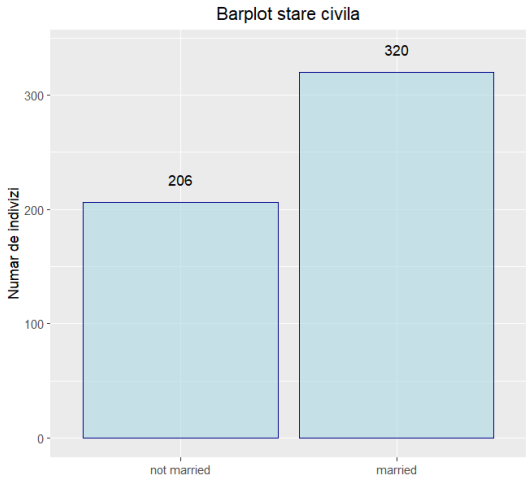
Graficul evidențiază distribuția eșantionului în funcție de rasa indivizilor. Majoritatea respondenților sunt de rasă albă (472), în timp ce minoritățile rasiale însumează doar 54 de cazuri. Această disproporție semnificativă indică un dezechilibru în structura eșantionului, care poate influența interpretarea modelelor predictive și generalizarea rezultatelor. Prin urmare, în etapele ulterioare ale analizei, va fi important să se țină cont de acest dezechilibru.

Graficul 3.12



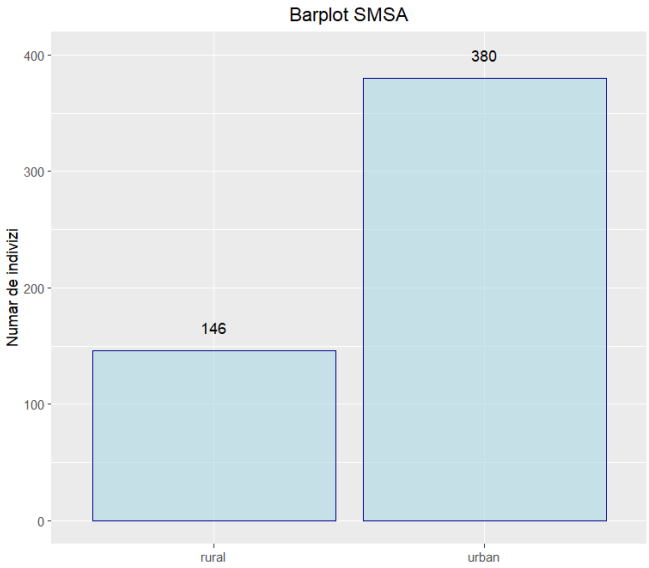
Graficul care reflectă distribuția pe sexe indică o repartizare echilibrată între bărbați (274) și femei (252), cu o ușoară predominanță masculină. Această uniformitate relativă sugerează că variabila sex nu suferă de dezechilibru de clasă major și poate fi utilizată în modelele predictive fără a necesita corecții suplimentare. Totodată, permite o analiză comparativă relevantă între sexe în ceea ce privește relațiile cu alte variabile, precum venitul, ocupația sau nivelul de educație.

Graficul 3.13



Graficul privind mediul de proveniență arată o predominanță clară a indivizilor din mediul urban (380), comparativ cu cei din mediul rural (146). Această distribuție indică o concentrare a populației analizate în zonele urbane, aspect ce poate influența semnificativ variabilele socio-economice precum accesul la educație, tipul ocupației sau nivelul de venit. Astfel, mediul de proveniență reprezintă o variabilă importantă în construirea și interpretarea modelelor predictive

Graficul 3.14



Graficul care prezintă distribuția indivizilor în funcție de starea civilă evidențiază o predominanță a persoanelor căsătorite (320), comparativ cu cele necăsătorite (206). Această diferență poate reflecta caracteristici socio-demografice ale populației analizate și are potențial explicativ în modelele predictive, mai ales în raport cu variabile precum venitul.

3.2. Identificarea și tratarea outlierilor și a valorilor lipsă

Outputul 3.1

```
> sum(is.na(wage1))
[1] 0
```

Valorile lipsă sunt un impediment în analiza datelor care, în cazul multor seturi de date, nu poate fi neglijat. La nivelul setului de date ce stă la baza acestui proiect totuși, nu există valori lipsă.

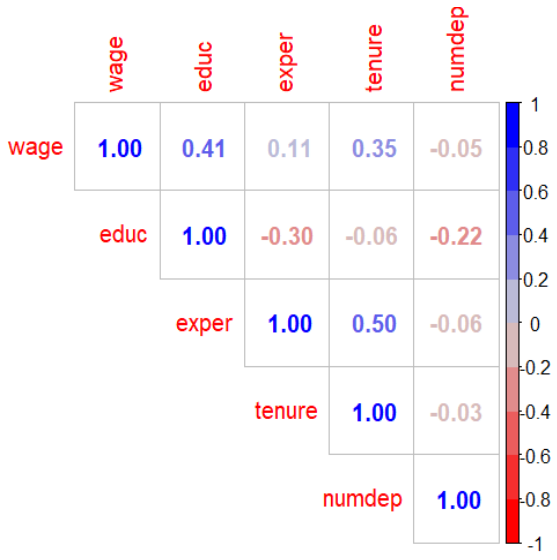
Outlierii, sau valorile extreme, pot afecta anumite distribuții prin prisma degradării importanței și eficacității mediei ca instrument de analizare a acelei variabile. Deși setul de date se confruntă cu probleme legate de outlieri pentru 3 din 5 variabile numerice, inclusiv variabila obiectiv, valorile nu sunt eronate și ilustrează caracteristici reale ale unor instanțe, motiv pentru care am ales să nu eliminăm aceste valori ci să le gestionăm în cadrul procesului de discretizare în momentul definirii intervalelor.

4. Procesarea datelor pentru modelare

Procesarea datelor reprezintă o etapă fundamentală în pregătirea setului de date pentru activitățile de modelare predictivă. Aceasta presupune transformarea și adaptarea variabilelor astfel încât să fie compatibile cu cerințele algoritmilor utilizați, menținând în același timp relevanța și integritatea informațională. În această secțiune sunt abordate operațiuni precum selecția variabilelor relevante, discretizarea variabilelor continue și recodificarea celor categoriale

4.1. Selectarea variabilelor numerice potrivite pentru modelare

Grafic 4.1: Matricea de corelație



Analizând matricea, variabila salariu prezintă o corelație pozitivă moderată cu educația (0.41) și tenure-ului (0.35), sugerând că niveluri mai ridicate de educație și experiență sunt asociate cu salarii mai mari. O examinare suplimentară dezvăluie interrelații între celelalte variabile. Educația prezintă o corelație negativă cu experiența (-0.30) sugerând că niveluri mai ridicate de educație pot fi asociate

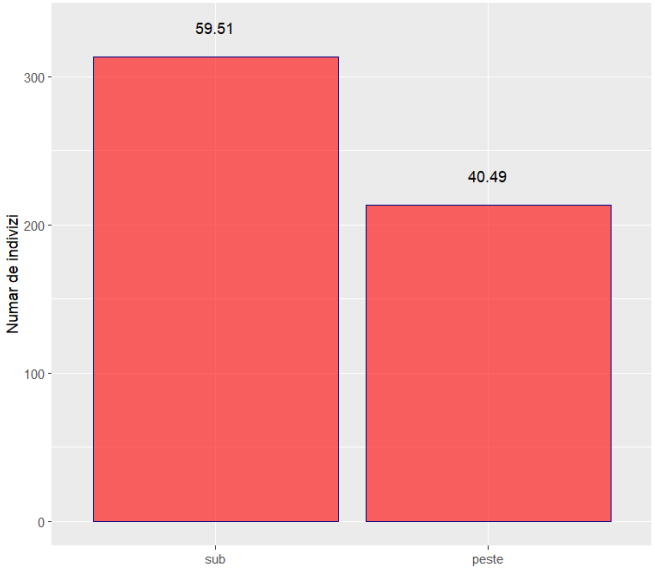
cu mai puțină experiență. Numărul de dependenți exercită o influență neglijabilă asupra tuturor variabilelor numerice, singura relație remarcabilă este cea cu educația, sugerând că o creștere a nivelului de educație duce la scăderea numărului de dependenți.

Trecând la procesul de selecție a variabilelor relevante, **educația este potrivită pentru modelare** întrucât prezintă cea mai strânsă corelație cu variabila obiectiv. Între tenure și experiență este moderată, astfel existând posibilitatea ca includerea ambelor variabile să fie redundantă. Dintre cele două, **tenure va fi variabila inclusă** deoarece are o corelație mult mai puternică cu variabila obiectiv. Numărul de dependenți prezintă o corelație neglijabilă cu salariul. Acest tipar fiind de așteptat în cazul datelor de angajare întrucât numărul de dependenți afectează venitul net al unui individ, nu salariul, prin capacitatea sa de beneficiu fiscal (Guner et. al, 2014). Din aceste motive numdep va fi, de asemenea, exclus din model.

4.2. Discretizarea variabilelor continue

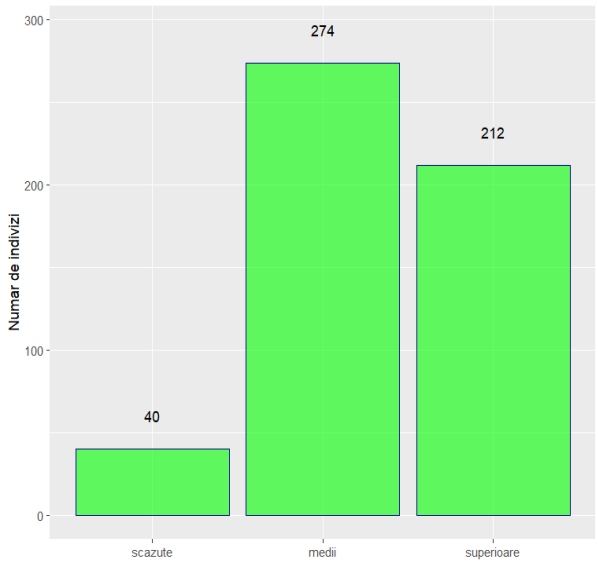
Procesul de discretizare are ca scop transformarea variabilelor continue, în variabile discrete pentru îndeplinii cerințele de procesare ale algorimilor de machine learning. Prima variabilă ce necesită discretizre este salariul. Pentru a crea arbori de decizie această variabilă va fi dihotomică. Am stabilit că punctul de tăiere ar trebui să fie 5.58 $/h care reprezintă salariul mediu din SUA anului 1976 conform „Soldajului asupra venitului național și conturile de produse, 1976-79” realizat de Biroului de Analiză Economică SUA.

Graficul 4.2: Frecvențele relativă a variabilei discretizate „wage”



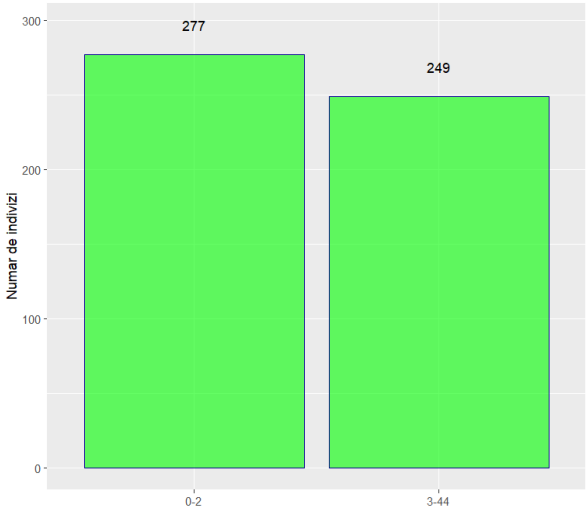
Această metodă de discretizare a dus la crearea unei nebalansări ușoare, 59.5% din indivizi având salarii sub salariul mediu iar 40.5% peste salariul mediu. Deși balansarea este observabilă aceasta nu atinge raportul de 2:1, fapt pentru care am decis că nu este necesară intervenția asupra distribuției.

Graficul 4.3: Frecvențele absolută a variabilei discretizate „educ”



Pentru discretizarea variabilei „educ” am selectat 3 intervale care conturează 3 categorii ierarhice ale nivelului de educație. De la 0 la 8 ani se consideră că cetățeanul are un nivel redus de studii, de la 9 la 12 ani individul are un nivell mediu de educație iar mai mult de 12 ani reprezintă studii superioare. Această structurare este dată de „Raportul ce vizează nivelul educațional în Statele Unite: 1977 și 1976” furnizat de Biroul de Recensământ al SUA. O abordare alternativă ar fi fost una în care studiile superioare sunt defalcate pe 4 cicluri: postliceal, licența, masteral și doctoral însă această abordare ar fi dus la crearea unor categorii cu puțini indivizi ca urmare a prezenței outlierilor.

Grafic 4.4: Frecvențele absolută a variabilei discretizate „tenure”



În literatura de specialitate nu poate fi identificată o metodă de discretizare a experienței pe contract nedeterminat . În aceste condiții am optat pentru algoritmul CAIM deoarece acesta se concentrează pe optimizarea discretizării variabilelor prin maximizarea legăturii cu atributul de clasă, asigurând în același timp un număr cât mai redus de intervale (Lavangnananda & Chattanachot, 2017). După discretizare au fost definite 2 clase, cei cu 0-2 ani pe contract nedeterminat si cei cu 3-44 ani pe contract nedeterminat. Câmpul mare de valori al celei de a doua grupe este dat de outlieri. Distribuția este una aproximativ egalitară cu 277 de oameni in prima grupa și cu 249 de oameni în a doua.Selectarea variabilelor categoriale potrivite pentru modelarePentru variabilele categoriale s-au calculat metrica Cramer’s V care indică intensitatea asocierii dintre categoriile variabilelor explicative și categoriile variabilei obiectiv. Cu cât valoarea este mai mare, cu atât categoriile

variabilei obiectiv sunt mai puternic asociate cu categoriile celorlalte variabile. Analizând valorile din Tabelul 4.1, variabila **tenure** înregistrează cel mai mare coeficient Cramer's V sugerând că vechimea are cea mai puternică asociere dintre toate variabilele considerate cu variabila dependentă, urmată de **econsect education si sex**. Pentru variabila smsa valoarea coeficientului scade cu 15%, delimitându-se astfel 4 variabile puternic asociate și trei slab asociate. Modelele vor fi antrenate luandu-se în considerare doar primele 4 variabile categoriale, variabilele **smsa, region și nonwhite fiind excluse din model**.

Tabel 4.1: Cramer's V pentru variabilele categorial

	Cramer's V
tenure	0.32710
econsect	0.32050
education	0.31860
sex	0.31050
smsa	0.16540
region	0.12180
nowhite	0.02382

Pentru o mai bună înțelegere asupra categoriilor variabilelor am decis să recodificăm categoriile pentru varaabila sex in „male" și „female"

5. Construirea si evaluarea modelelor CART, QUEST și CHAID

Modelele CART, QUEST și CHAID sunt utilizate frecvent în analiza predictivă pentru clasificarea observațiilor pe baza unui set de variabile explicative. CART construiește doar arbori binari, nodurile părinte având exact doi copii. Procesul de divizare utilizează criteriul Twoing, iar structura finală a arborelui este optimizată prin metoda de tăiere bazată pe complexitatea costului. Acest algoritm are implementat o metodă automată de discretizare motiv pentru care v-om folosii valorile continue pentru variabilele care inițial au avut acest tip, exceptând variabila obiectiv

Regula de divizare din cadrul algoritmului QUEST presupune inițial existența unei variabile țintă de tip continuu. Comparativ cu alte metode, QUEST oferă o viteză de calcul superioară și reduce riscul apariției bias-ului în selecția predictorilor. Este deosebit de eficient în tratarea variabilelor multicategoriale, deși funcționează exclusiv cu date binare. Particularitățile sale

constau în tehnicile utilizate pentru alegerea criteriilor de divizare atât în nodul rădăcină, cât și în nodurile descendente, precum și în numărul de ramificații generate la fiecare pas al construcției arborelui (Lin & Fan, 2019).

Algoritmul CHAID, dezvoltat de Kass (1980), utilizează testul chi-pătrat (χ^2) pentru a stabili regulile de divizare în cadrul arborelui decizional. Acest test măsoară gradul de asociere dintre variabile, iar o valoare ridicată a statisticii χ^2 indică o dependență semnificativă și, implicit, o probabilitate crescută ca variabila să fie relevantă în procesul de predicție. Algoritmul folosește valorile p pentru a decide dacă divizarea unui nod trebuie continuată, evaluând succesiv toate variabilele candidate. Pentru fiecare variabilă, se testează semnificația diferențelor dintre categoriile variabilei dependente. Categoriile considerate nesemnificative sunt reunite într-un grup omogen, iar procesul de testare și comasare continuă până când nu mai există diferențe statistic semnificative. Acest algoritm nu necesită un criteriu de fasonare (Lin & Fan, 2019).

5.1. Împărțirea setului de date în subseturi de antrenament și testare

Seturile de train și de test au fost create cu ajutorul pachetului „caret”. Testul de antrenament conține 70% din date iar cel de test restul de 30%, nu a fost utilizata nici o metodă de reeșantionare sau validare încrucișată. De asemenea, toate modelele prezentate anterior au fost antrenate și testatepe aceleași seturi.

5.2. Parametrii de creștere

Parametrii de creștere sunt folosiți pentru a se stabilii limitele de divizare sau de fasonare ori sunt determinate alte aspecte ce vizează dimensiunile arborelui astfel încât să se evite overfittingul. Niciunul dintre cei 3 algoritmi nu a trecut printr-un proces de tuning, criteriile de creștere rămânând la valorile standard.Pentru algoritmul CART se pot stabilii 9 parametrii de creștere prezentați în Tabelul 5.1. Aceștia sunt:

- **Minsplit**: Numărul minim de observații necesare într-un nod pentru ca acesta să poată fi divizat. Dacă un nod conține mai puțin de 20 de observații, nu va fi luată în considerare divizarea lui.
- **Minbucket:** Numărul minim de observații care trebuie să existe într-un nod terminal (frunză). Asigură că arborele nu produce noduri terminale cu foarte puține date, prevenind astfel supraînvățarea (overfitting).
- **Cp:** Parametru de complexitate controlează cât de mult trebuie să scadă eroarea pentru a justifica o nouă divizare. Cu cât valoarea este mai mică, cu atât arborele poate deveni mai complex. Un cp de 0.01 indică o echilibrare între performanță și complexitate.

- **Maxcompete:** Numărul maxim de alternative (splituri competitoare) care sunt păstrate pentru fiecare nod, în scopuri de analiză comparativă, dar care nu sunt utilizate în mod direct în modelul final.
- **Maxsurrogate:** Numărul maxim de variabile surogat folosite atunci când datele lipsesc pentru variabila principală utilizată la o divizare. Aceste variabile încearcă să mimeze comportamentul variabilei principale.
- **Usesurrogate:** Indică modul în care sunt utilizate variabilele surogat: 2 înseamnă că sunt folosite pentru clasificare atunci când lipsesc valorile variabilei principale.
- **Surrogatestyle:** Determină stilul de selecție a surogatelor. 0 indică stilul complet bazat pe acuratețea surogatelor în replicarea divizării originale.
- **Maxdepth:** Adâncimea maximă a arborelui, adică numărul maxim de niveluri pe care le poate atinge. O valoare mare, cum este 30, permite un arbore foarte detaliat, dar poate crește riscul de overfitting.
- **Xval:** Numărul de folduri utilizate în validarea încrucișată (cross-validation) pentru estimarea erorii arborelui. Cu xval = 10, se aplică validare încrucișată în 10 folduri pentru evaluarea performanței.

Tabelul 5.1: Parametrii de creștere a modelului CART

Parametru	Valoare
minsplit	20
minbucket	7
cp	0.01
maxcompete	4
maxsurrogate	5
usesurrogate	2
surrogatestyle	0
maxdepth	30
xval	10

Pentru algoritmul QUEST are un total de 34 de criterii de creștere, cei mai importanți au fost prezentați în Tabelul 5.2, iar interpretările acestora sunt:

- **Nresample**: Numărul de resamplinguri (bootstrapuri) utilizate pentru testele statistice aplicate la alegerea variabilei de split. O valoare mare, cum este 9999, asigură o estimare stabilă a semnificației statistice, dar crește timpul de calcul.
- **Minsplit**: Numărul minim de observații necesare într-un nod pentru ca acesta să poată fi divizat. Dacă un nod conține mai puțin de 20 de observații, nu va fi luată în considerare divizarea lui.
- **Minbucket**: Numărul minim de observații care trebuie să existe într-un nod terminal (frunză). Asigură că arborele nu produce noduri terminale cu foarte puține date, prevenind astfel supra învățarea.
- **Minprob**: Probabilitatea minimă necesară ca o ramură să fie păstrată. O valoare de 0.01 înseamnă că nodurile terminale care conțin mai puțin de 1% din totalul observațiilor pot fi eliminate, pentru a reduce complexitatea modelului.
- **Maxvar**: Numărul maxim de variabile candidate evaluate pentru o posibilă divizare. Inf (infinit) înseamnă că toate variabilele disponibile vor fi luate în considerare, fără restricții.
- **Stump**: Indică dacă arborele este limitat la un singur split (un „ciot”). Valoarea FALSE înseamnă că arborele poate avea mai multe niveluri de ramificare, adică nu este limitat la un singur nivel.
- **Maxdepth**: Adâncimea maximă a arborelui, adică numărul maxim de niveluri pe care le poate atinge. O valoare Inf (nelimitată) permite arborelui să crească cât este necesar, dar poate duce la supra învățare dacă nu este controlată altfel.
- **Multiway**: Indică dacă sunt permise divizări cu mai mult de două ramuri (multiway splits). Valoarea FALSE specifică faptul că sunt utilizate doar divizări binare, ceea ce este în acord cu principiile clasice ale algoritmului QUEST.

Tabelul 5.2: Parametrii de creștere a modelului QUEST

Parametru	Valoare
nresample	9999
minsplit	20
minbucket	7
minprob	0.01
maxvar	Inf
stump	FALSE
maxdepth	Inf
multiway	FALSE

Algoritmul CHAID are doar 8 parametrii inițiali ce pot fi ajustați, iar interpretările informațiilor din din Tabelul 5.3 sunt:

- **alpha2**: Reprezintă nivelul de semnificație utilizat pentru testul de fuziune a categoriilor în cadrul nodurilor. Un prag de 0.05 indică faptul că două categorii vor fi comasate doar dacă nu există diferențe semnificative statistic între ele ($p > 0.05$), reducând complexitatea arborelui.
- **alpha3**: În mod standard, acest parametru controlează semnificația pentru pre-alocarea variabilelor în noduri copil, însă valoarea -1 indică faptul că această verificare este dezactivată, toate variabilele fiind eligibile pentru evaluare fără filtrare prealabilă.
- **alpha4**: Este nivelul de semnificație folosit pentru divizarea nodurilor. Cu un prag de 0.05, nodul curent va fi divizat doar dacă se detectează o relație semnificativă între variabila dependentă și predictori (conform testului chi-pătrat).
- **minsplit**: Indică numărul minim de observații necesare într-un nod pentru ca acesta să fie luat în considerare pentru divizare. Previne divizările pe baza unui volum de date prea redus.
- **minbucket**: Reprezintă numărul minim de observații într-un nod terminal. Asigură faptul că frunzele arborelui nu conțin prea puține observații, reducând astfel riscul de supra ajustare.
- **minprob**: Este proporția minimă a unui nod față de totalul eșantionului. Un nod trebuie să conțină cel puțin 1% din totalul observațiilor pentru a fi eligibil pentru divizare, evitând astfel generarea de ramuri nesemnificative.
- **stump**: Specifică faptul că arborele nu este limitat la un singur nivel (nu este un "stump"). Prin urmare, se permite construirea unui arbore complet, cu ramificări multiple.

- **maxheight**: O valoare negativă semnalează că nu există o limitare predefinită a înălțimii arborelui, permițând algoritmului să continue divizarea până când toate condițiile de oprire sunt îndeplinite (conform testelor de semnificație și limită de observații).

Tabelul 5.3: Parametrii de creștere a modelului CHAID

Parametru	Valoare
alpha2	0.05
alpha3	-1.00
alpha4	0.05
minsplit	20.00
minbucket	7.00
minprob	0.01
stump	0.00
maxheight	-1.00

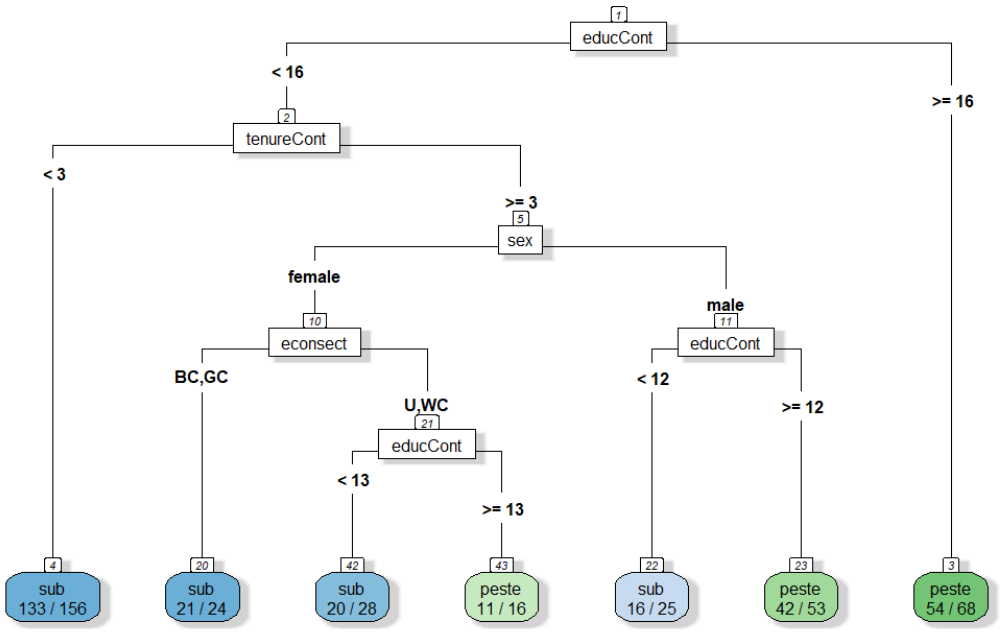
5.3. Reprezentarea grafică a modelelor

Reprezentările grafice sunt un mod ideal de a determina interpretabilitatea, importanța variabilelor în model și relațiile variabilelor din model. Figurile propriu zise rezultate variază, ca stil, în funcție de model deoarece acestea au fost construite cu ajutorul a 3 pachete separate. În Figura 5.1 este prezentat fluxul prin care o instanță v-a trece pentru a fi clasificată. Rădăcina arborelui este reprezentată de variabila „educCont" (educație continuă), care apare în poziția cea mai înaltă, ceea ce indică faptul că aceasta are cea mai mare importanță în procesul de clasificare. În mod concret, prima împărțire se realizează pe baza valorii prag de 16 a acestei variabile. Astfel, observațiile se separă în două ramuri distincte: cele cu educCont < 16 sunt direcționate spre stânga, iar cele cu educCont ≥ 16 sunt direcționate spre dreapta.

Ramura stângă a arborelui este ulterior segmentată în funcție de variabila „tenureCont", cu un prag de 3. Dacă tenureCont < 3, atunci se ajunge la un nod terminal în care majoritatea observațiilor sunt clasificate ca „sub" (133 din 156). În schimb, dacă tenureCont ≥ 3, urmează o împărțire pe baza variabilei „sex", distingând între femei și bărbați. Pentru femei, împărțirea se continuă prin variabila „econsect" (sectoare economice), cu două ramuri: BC, GC și U, WC. În ramura BC, GC, rezultatul este predominant „sub", în timp ce în ramura U, WC, clasificarea se

face în funcție de educCont, cu un nou prag de 13. Astfel, pentru femeile din sectoarele U și WC cu educCont ≥ 13, clasificarea se schimbă în „peste”, ceea ce sugerează că educația suplimentară compensează influența negativă a sectorului de muncă. Pe ramura corespunzătoare bărbaților (sex = male), decizia este din nou ghidată de variabila educCont, dar de această dată cu un prag de 12. Cei cu educație sub acest prag sunt majoritar clasificați ca „sub”, în timp ce bărbații cu educație continuă ≥ 12 sunt în general încadrați în categoria „peste”. Această diviziune arată o relație consistentă între nivelul educației și probabilitatea de a obține un rezultat favorabil (peste), mai ales în rândul bărbaților. Prin urmare, modelul subliniază influența dominantă a variabilei educCont, urmată de tenureCont, sex, econsect

Figura 5.1: Reprezentarea grafică a modelului CART

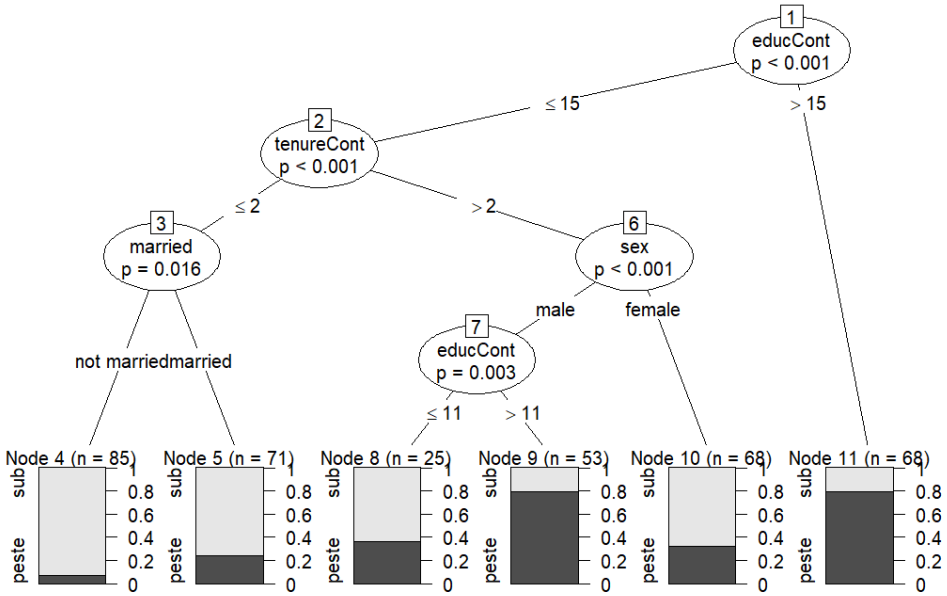


Pentru algoritmul QUEST, reprezentarea din Figura 5.2, schema debutează cu variabila „educCont”, aleasă ca prim criteriu de divizare cu un nivel de semnificație ridicat ($p < 0.001$). Împărțirea se face în funcție de pragul 15, ceea ce sugerează că acest nivel de educație este esențial în determinarea categoriei rezultate. Observațiile cu educCont > 15 sunt direcționate către nodul terminal 11, în care se remarcă o proporție ridicată de rezultate „peste”, indicând o corelație pozitivă între educația ridicată și obținerea unui rezultat favorabil.

Pentru observațiile cu educCont ≤ 15, se continuă segmentarea cu variabila „tenureCont”, utilizată în nodul 2 ($p < 0.001$), cu un prag de 2. Persoanele cu vechime ≤ 2 sunt ulterior clasificate în funcție de statutul marital („married”), variabilă introdusă în nodul 3 ($p = 0.016$). Astfel, indivizii necăsătoriți sunt majoritar clasificați în categoria „sub” (nodul 4), în timp ce cei căsătoriți (nodul 5) prezintă o distribuție mai echilibrată, dar tot cu o predominanță a categoriei „sub”. Această ramură subliniază importanța cumulativă a educației, vechimii și statutului marital în predicția rezultatului negativ.

În partea cealaltă a arborelui, unde tenureCont > 2, segmentarea se face în funcție de sex (nodul 6, $p < 0.001$). Femeile sunt clasificate direct în nodul 10, unde proporția de rezultate „sub” rămâne ridicată. În cazul bărbaților, segmentarea este rafinată printr-un al doilea prag pentru educCont (nodul 7, $p = 0.003$), de data aceasta cu un prag de 11. Cei cu educație ≤ 11 sunt incluși în nodul 8, dominat de rezultatul „sub”, în timp ce bărbații cu educație > 11 (nodul 9) sunt mai bine reprezentați în categoria „peste”, indicând o influență pozitivă a unui nivel educațional superior, chiar și într-un context în care celelalte caracteristici nu sunt favorabile. Această structură arboreală relevă o serie de interacțiuni complexe între variabilele explicative, în care educația joacă un rol central, urmată de vechime, sex și statut marital

Figura 5.2: Reprezentarea grafică a modelului QUEST

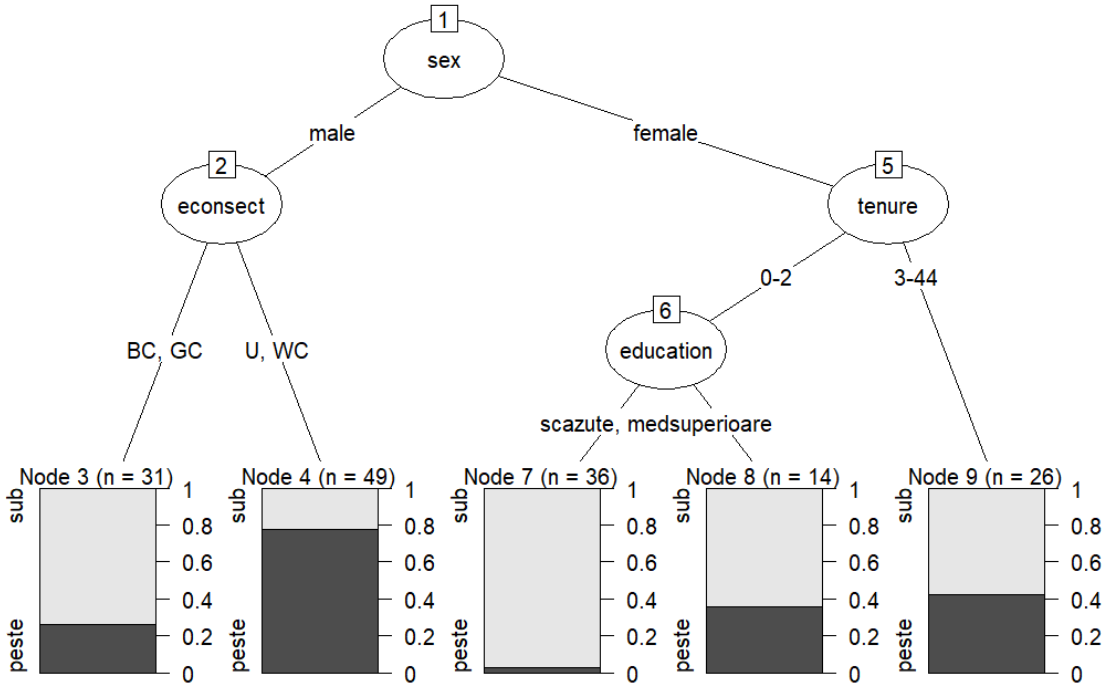


În figura 5.3 se poate observa arhitectura modelului CHAID. La nivelul rădăcinii arborelui (nodul 1), variabila de segmentare inițială este „sex”, ceea ce sugerează că genul reprezintă factorul cel mai semnificativ din punct de vedere statistic în explicarea variabilei țintă. Modelul separă populația în două ramuri distincte: bărbați și femei. Această decizie inițială evidențiază o relație diferențiată între sexe în ceea ce privește determinarea salariului.

Pentru bărbați, diviziunea se realizează pe baza variabilei „econsect”, care este împărțit în două categorii: BC și GC, respectiv U și WC. Observațiile din sectoarele BC și GC sunt alocate nodului 3, unde predomină rezultatul „sub”, cu o proporție semnificativă de cazuri peste, însă minoritară. În schimb, bărbații din sectoarele U și WC (nodul 4) au o distribuție inversată, unde majoritatea sunt clasificați ca „peste”. Această diferență sugerează că pentru bărbați, sectorul de activitate influențează în mod direct probabilitatea de succes, în defavoarea celor din sectoarele considerate mai stabile sau bine remunerate.

Pentru femei, segmentarea continuă pe baza vechimii în muncă, prin variabila „tenure” (nodul 5), care este împărțită între intervalele 0–2 și 3–44 ani. Această divizare evidențiază faptul că tenure-ul este un determinant semnificativ pentru femei. Femeile cu o vechime profesională pe contract nedeterminat mică (0–2 ani) sunt mai departe împărțite în funcție de nivelul educațional (nodul 6), diferențiat între „scăzute, medii” și „superioare”. Cele cu educație scăzută si medii (nodul 7) sunt aproape integral clasificate ca „sub”, în timp ce femeile cu educație superioara(nodul 8) au o distribuție echilibrată, deși cu o ușoară predominanță pentru „sub”. Femeile cu o vechime profesională mai mare (nodul 9), indiferent de nivelul educațional, prezintă o tendință ușor favorabilă către rezultatul „peste”, ceea ce confirmă faptul că experiența compensează în parte nivelul educației scăzute. Prin urmare, modelul CHAID identifică trei variabile esențiale în procesul de decizie: sexul, sectorul economic și vechimea profesională, cu influență suplimentară din partea educației

Figura 5.3: Reprezentarea grafică a modelului CHAID



5.4. Prezentarea comparativă a performanțelor modelelor

În această secțiune, se realizează o analiză comparativă a performanțelor modelelor dezvoltate, utilizând indicatori de evaluare relevanți precum acuratețea, sensibilitatea, coeficientul Kappa și scorul F1. Compararea rezultatelor obținute permite identificarea modelului care oferă cele mai bune rezultate în contextul problemei studiate, evidențiind punctele forte și limitările fiecăruia

Tabelul 5.4: Metricile de performanță a celor 3 modele, afișate comparativ

	Train				Test			
	Acc(%)	**Recall(%)**	**F1**	**Kappa**	**Acc(%)**	**Recall(%)**	**F1**	**Kappa**
CART	80.27	86.36	0.83	0.58	74.36	80.65	0.78	0.46
QUEST	78.65	88.64	0.83	0.54	75.64	84.95	0.80	0.48
CHAID	68.92	80.45	0.75	0.33	76.92	88.17	0.82	0.50

Pe setul de antrenare, modelul CART înregistrează cea mai mare acuratețe (80,27%), urmat de QUEST (78,65%) și CHAID (68,92%). Aceasta indică o potrivire superioară a modelului CART pe datele de antrenare, în timp ce CHAID prezintă performanțe mai slabe. În ceea ce privește recall-ul, care măsoară proporția de salarii sub medie identificate corect, QUEST obține cel mai bun rezultat (88,64%), urmat de CART (86,36%) și CHAID (80,45%). Scorul F1, care reflectă echilibrul între precizie și recall, este identic pentru CART și QUEST (0,83), sugerând o performanță comparabilă în detectarea salariilor sub medie, în timp ce CHAID (0,75) este inferior. Coeficientul Kappa, ce evaluează acordul dintre predicții și valorile reale, ajustat pentru clasificarea aleatoare, este cel mai ridicat pentru CART (0,58), urmat de QUEST (0,54) și CHAID (0,33), confirmând superioritatea CART pe setul de antrenare.

Pe setul de testare, care reflectă capacitatea de generalizare, CHAID obține cea mai mare acuratețe (76,92%), urmat de QUEST (75,64%) și CART (74,36%). Această inversare față de performanța pe antrenare sugerează că CHAID generalizează mai bine pe date nevăzute. Pentru recall, CHAID conduce cu 88,17%, urmat de QUEST (84,95%) și CART (80,65%), evidențiind capacitatea superioară a CHAID de a identifica salariile sub medie pe setul de testare. Scorul F1 urmează același tipar, CHAID obținând 0,82, urmat de QUEST (0,80) și CART (0,78). Coeficientul Kappa pe testare este cel mai ridicat pentru CHAID (0,50), urmat de QUEST (0,48) și CART (0,46), ceea ce subliniază consistența CHAID în predicții pe date noi.

Comparativ, CART tinde să supra adapteze datele de antrenare, având cele mai bune valori pentru acuratețe și Kappa pe acest set, dar performanțe mai slabe pe testare, ceea ce sugerează o complexitate excesivă a modelului. În contrast, CHAID prezintă performanțe superioare pe setul de test comparative cu cel de antrenament, fapt ce poate fi explicat de lipsa prezentei unui system de fasonare ce implică cross validation, o diferență în distribuția datelor de test si de antrenament sau underfittingul. Modelul QUEST este cel mai echilibrat dintre cele 3, diferența de acuratețe dintre setul de test și cel de antrenament fiind de doar 3% ceea ce înseamnă că acestea are o capacitate ridicată de generalizare.

Concluzii

Studiul de față, axat pe clasificarea salariului în funcție de caracteristicile socio-demografice ale indivizilor prin intermediul arborilor de decizie, a evidențiat o serie de relații relevante între factorii personali și profesionali și nivelul salarial, utilizând datele din Current Population Survey (1976). Prin aplicarea algoritmilor CART, QUEST și CHAID, precum și a unei metodologii riguroase de procesare și analiză, cercetarea a reușit să identifice tipare semnificative privind influența educației, experienței, genului și sectorului economic asupra veniturilor.

Unul dintre principalele rezultate ale analizei este **rolul esențial al educației ca determinant al salariului**, aceasta fiind **variabila de segmentare inițială în modelele CART și QUEST**. Aceste modele arată că un nivel educațional mai ridicat este asociat cu o probabilitate crescută de a obține un salariu peste medie. Modelul CHAID, în schimb, aduce în prim-plan **genul ca factor decizional principal**, relevând diferențe semnificative între bărbați și femei în ceea ce privește influențele asupra veniturilor. În acest context, sectorul **economic se dovedește esențial pentru bărbați, în timp ce pentru femei, experiența profesională și educația joacă un rol mai important**, subliniind astfel existența unor disparități de gen în structura salarială.

Vechimea în muncă, analizată prin prisma variabilei tenure, este de asemenea un factor important, cu praguri distincte ce separă categoriile salariale. Modelele **CART și QUEST confirmă influența tenure-ului** asupra salariului, în timp ce **CHAID accentuează rolul său compensator pentru femeile cu nivel educațional redus**. Această corelație indică faptul că **stabilitatea la locul de muncă și experiența acumulată contribuie la creșterea veniturilor.**

Compararea performanțelor algoritmilor aplicați a evidențiat diferențe relevante în ceea ce privește capacitatea lor de generalizare. **CART, deși performant pe setul de antrenament, a suferit o scădere notabilă pe setul de testare**, **semnalând o posibilă supra adaptare**. CHAID s-a remarcat printr-o performanță superioară pe datele de test, diferența de acuratețe fiind de **9% lucru ce poate sugera underfitting sau o împărțire defectuoasă**, în timp ce **QUEST a oferit un echilibru între stabilitate și precizie**. Interpretabilitatea este unul din avantajele folosirii arborilor de decizie, acest criteriu fiind îndeplinit de toate modelele, **acestea având între 4 și 6 nivele**.

Din punct de vedere metodologic, **discretizarea variabilelor continue**, precum educația și salariul, **a fost fundamentată teoretic și susținută prin referințe la surse oficiale**. Totuși, **decizia de a păstra outlierii prin discretizare**, în special în cazul tenure-ului, **poate introduce zgomot în date și influența precizia modelelor**. Acest aspect indică oportunitatea unor analize suplimentare care să evalueze impactul acestor valori extreme asupra performanței predictive.

Bibliografie

Bureau, U. S. (1977). *Educational Attainment in the United States: 1977 & 1976.*

Guner, N., Kaygusuz, R., & Ventura, G. (2014). Income taxation of U.S. households: Facts and parametric estimates. *Review of Economic Dynamics*, 17(4), 559-581.

Lavangnananda, K., & Chattanachot, S. (2017). Study of discretization methods in classification. *ResearchGate*, 50-55.

Lin, C., & Fan, C. (2019). Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. *Journal of Asian Architecture and Building Engineering*, 18(6), 539–553.

Missouri, U. o. (1976). *Prices and Wages by Decade: 1970-1979.*

Urgan, S., & Küsbeci, P. (2022). A Conceptual Study on Collar Classification in Employees. *7th INTERNATIONAL NEW YORK CONFERENCE ON EVOLVING TRENDS IN INTERDISCIPLINARY RESEARCH & PRACTICES*, (pg. 187-192). New York.