O datasetu:

Dataset koji je korišćen služi za predikciju ponovnog pojavljivanja raka dojki kod prethodno obolelih osoba. Sadrži 9 nezavisnih atributa i 1 zavisan koji predstavlja klasu koja treba da se predvidi. Sadrži 285 instanci i 10 kolona. Atributi su sledeći:

Age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99

Menopause: lt40, ge40, premeno

Tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59

Inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39

Node-caps: yes, no

Deg-malig: 1, 2, 3

Breast: left, right

Breast-quad: left-up, left-low, right-up, right-low

Irradiat: yes, no

Class: no-reccurence-events, reccurence -events

O radu:

Model stabla odlučivanja je pravljen u Jupyter Notebook-u, bez pomoći biblioteka. Izvršena je analiza nad dataset-om i utvrđeno je da postoje duplikati, koji su odmah potom uklonjeni. Nema nedostajućih vrednosti i podaci su dobro balansirani, nije bilo potrebe za dodatnim prečišćavanjem podataka. Svi atributi su kategorički, nije neophodno vršiti normalizaciju ili standardizaciju.

Kreirane su pomoćne funkcije za implementaciju algoritma : gini\_impurity() za računanje GINI indeksa čvora, gini\_total() za računanje GINI\_total podele na decu čvorove i compute\_accuracy() za računanje preciznosti klasifikacije.

Kreirana je klasa Node koja predstavlja čvor stabla kome su prirduženi atributi: left (levi čvor dete), right (desni čvor dete), feature\_question (feature pitanja po kom se deli čvor) i question (vrednost (skup vrednosti) tog feature-a), class\_val (vrednost klasnog atributa u tom čvoru).

Kreirana je klasa DecisionTree koja implementira algoritam koja sadrži: root (koren stabla, poziva funkciju koja kreira stablo odluke), features, target, class\_col (ime klase koju predviđamo). Funkcije koje implementira su: all\_attributes\_equal() koja vraća True ukoliko svi slogovi imaju jednake vrednosti atributa (prvi uslov za kreiranje terminalnog čvora), homogenous() koja vraća True ukoliko su slogovi homogeni, most\_valuable\_split() koja vraća feature i deo njegovog skupa vrednosti za koji je dobit podele najveći, decision\_tree\_classifier() koji kreira drvo i fit() koji dodeljuje to drvo root čvoru stabla.

Algoritam za kreiranje stabla(funkcija decision\_tree\_classifier()) koristi most\_valuable\_split da odredi pitanje koje nudi najveću dobit. Nakon toga, filtriraju se podaci na one koji pripadaju top skupu vrednosti feature-a (data\_right) i na one koji ne pripadaju (data\_left) i rekurzivno se poziva funkcija decision\_tree\_classifier() za svaki od njih sve dok se ne dođe do uslova za zaustavljanje (svi slogovi su homogeni, vrednosti atributa u svim slogovima ista ili premali broj instanci (preprunning)), nakon čega se kreira i vraća terminalni čvor sa odgovarajućom vrednošću za klasu koja se predviđa.

Bez ikakve validacije, vraćen je score 0.88, pri čemu je dodat dodatni uslov za preprunning stabla (ukoliko je broj instanci manji od zadatog broja (optimalno je 3 u ovom slučaju), kreirati terminalni čvor).

Uz 5 Fold cross validaciju (ručno implementiranu) vraćeni su score-vi od 0.65 do 0.68 uz isprobavanje različitih parametara za odsecanje stabla. Matrice konfuzije pokazuju da se pozitivne vrednosti većinom dobro klasifikuju, dok se negativne vrednosti klasifikuju gore.

Kod:

Poslat je kod u Notepad-u zajedno sa Jupyter notebook fajlom (sa naznačenim obavljenim koracima) koji je pokrenut uz prikazane rezultate.