Proiect AI : Machine learning on Post-Operative Data using MLP

Student : Matei Obogeanu

Data : 6 mai 2024

Introducere

Proiectul de față propune rezolvarea unei probleme de clasificare pentru determinarea următoarei etape de clasare a pacienților aflați la recuperare după operație. Atributele pe baza cărora se face această clasificare se referă în marea majoritate temperatura măsurată a corpului.

Baza de date conține **90** de înregistrări cu câte **8** atribute (features), 7 date categoriale, una numerică și **o** clasă de obiectiv (target) :

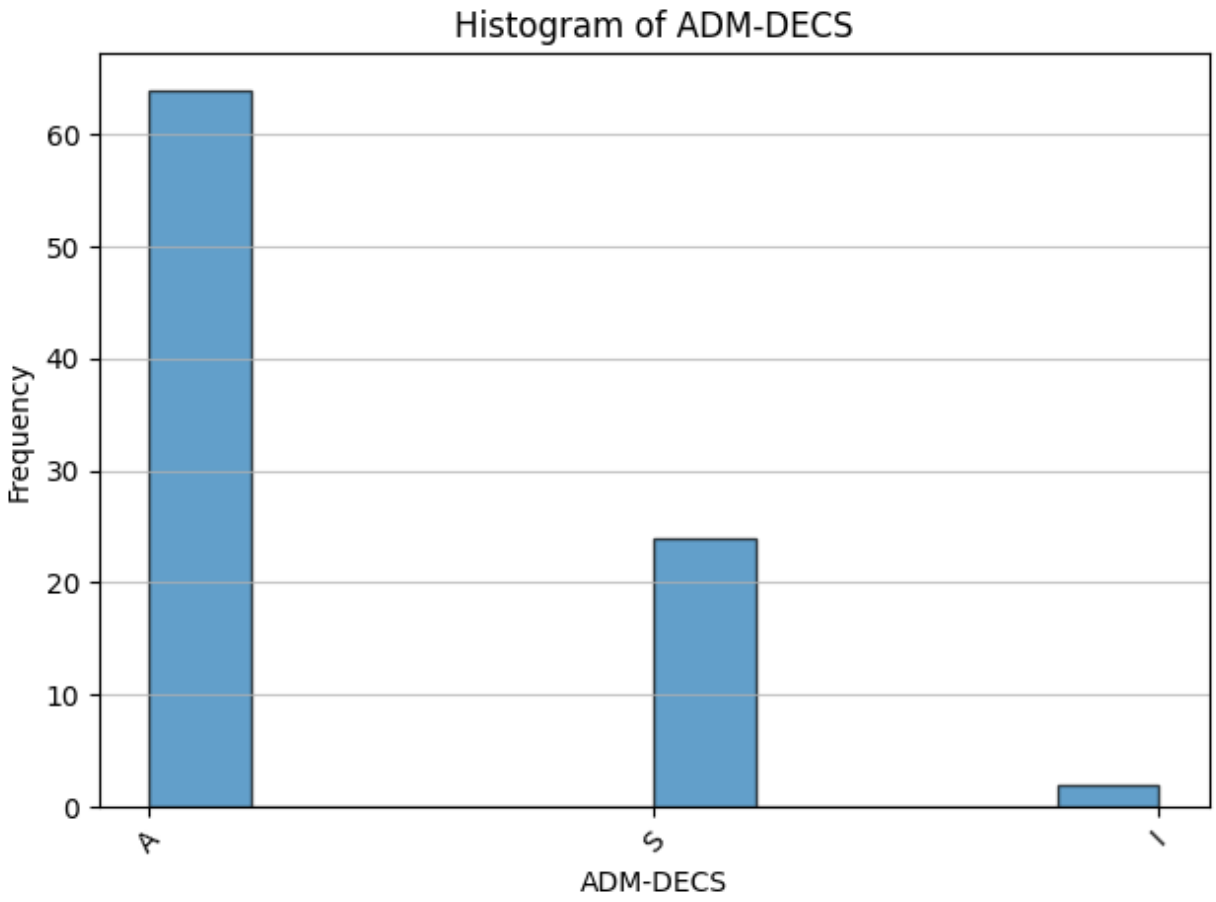
1. L-CORE (patient's internal temperature in C): high (> 37), mid (>= 36 and <= 37), low (< 36)
2. L-SURF (patient's surface temperature in C): high (> 36.5), mid (>= 36.5 and <= 35), low (< 35)
3. L-O2 (oxygen saturation in %):excellent (>= 98), good (>= 90 and < 98), fair (>= 80 and < 90), poor (< 80)
4. L-BP (last measurement of blood pressure): high (> 130/90), mid (<= 130/90 and >= 90/70), low (< 90/70)
5. SURF-STBL (stability of patient's surface temperature): stable, mod-stable, unstable
6. CORE-STBL (stability of patient's core temperature) : stable, mod-stable, unstable
7. BP-STBL (stability of patient's blood pressure) Șstable, mod-stable, unstable
8. COMFORT * (patient's perceived comfort at discharge, measured as an integer between 0 and 20)
9. decision ADM-DECS (discharge decision): I (patient sent to Intensive Care Unit),S (patient prepared to go home), A (patient sent to general hospital floor)

Librării utilizate:

1. **ucimlrpo** pentru încărcarea bazei de date
2. **sklearn.preprocessing** pentru codarea informațiilor
3. **sklearn.neural_network** pentru clasificatorul MLP
4. **sklearn.model_selection** pentru cross validation pe setul întreg de date

Preprocesarea Datelor

*În baza de date există înregistrări pentru care nu avem date pentru atributul COMFORT. Astfel, deoarece doar pentru 3/90 înregistrări lipsesc date, am ales să completez cu media gradului de confort al restului de înregistrări. O alternativă ar fi fost ștergerea acestor înregistrări și aplicarea clasificatorului pe restul de date, dar în acest caz m-aș fi lovit peste o altă problemă: baza de date este dezechilibrată conținând doar două înregistrări pentru elementele etichetate cu 'I' (pacient trimis la terapie intensivă). Odată cu eliminarea uneia dintre cele două înregistrări, performanța antrenării modelului m-aș fi așteptat să scadă foarte mult (sub pragul de acuratețe pe care îl promite sursa bazei de date, de 48%)



Am calculat de asemenea și frecvența relativă de apariție a fiecărei etichete, obținând aproximativ:

- 71% pentru clasa 'A'
- 27% pentru clasa 'S'
- 2% pentru clasa 'I' – contribuie mult la dezechilibrarea bazei de date

Codarea etichetelor și a atributelor categoriale am realizat-o cu un LabelEncoder din librăria sklearn.preproccessing, iar nivelului de confort (singurul atribut **numeric**) i am aplicat o scalare standard cu un StandardScaler din aceeași librărie.

Antrenarea modelului am realizat-o cu un clasificator MLP căruia i s-au variat următorul set de hyper-parametrii:

- learning rate : 0.1 sau 0.01
- hidden layers : (100, 200), (100, 100), (100, 50)

Am variat **bazele de antrenare / testare** prin cross-validation cu factor de 1/4 cu ajutorul cross_val_score din sklearn.model_selection. Acuratețea per eșantion de validare este evaluată după numărul de predicții corecte raportat la dimensiunea eșantionului, iar la final se mediază toate cele 4 validări pentru obținerea acurateței medii finale pentru hyper-parametrii aleși:

Hidden Layer 1	Hidden Layer 2	Learning Rate	Acccuracy
100	200	0.1	0.6240118577075098
100	200	0.01	0.5553359683794467
100	100	0.1	0.566699604743083
100	100	0.01	0.5345849802371542
100	50	0.1	0.5785573122529644
100	50	0.01	0.5785573122529644

O acuratețe bună ar trebui să se apropie riguros de 1, dar în acest caz, din cauza bazei de date dezechilibrate se atinge un maxim de aproximativ 62% pentru rețeaua neuronală cu cei mai mulți perceptroni pe cel două straturi și learning rate-ul cel mai mare.