Системи одлучивања у медицини (13E053COM)
Пројектни задатак 2024/25

Пројекат се може израдити индивидуално или у пару, брани се у било ком испитном року и учествује са 30% укупне оцене. Студент или група студената пријављује се за једну од база података наведених у табели путем тима на *MS Teams* платформи. Избор је могућ искључиво међу базама које до тог тренутка нису већ резервисане од стране друге групе.

Финални извештај треба да садржи све релевантне информације о изабраној бази података, као и резултате и коментаре за појединачне елементе анализе.

Анализа скупа података [10]

Потребно је извршити детаљну анализу скупа података. Уколико у датом скупу постоје недостајуће или неадекватне вредности, неопходно је заменити их очекиваним вредностима или уклонити, уз одговарајуће образложење избора приступа.

За све номиналне или категоријске атрибуте потребно је применити одговарајуће поступке кодирања и превести их у нумеричке атрибуте, уз образложење употребљене методе.

Класе чија учешће у укупној популацији износи мање од одабране границе (нпр. 3–5%) сматрају се малобројним. Све такве класе треба објединити у нову класу под називом „Остало“.

Селекција обележја [13]

Потребно је израчунати информациону добит за свако обележје и прокоментарисати добијене резултате. На основу тих резултата одабрати 10 најрепрезентативнијих обележја (ако скуп садржи мање од 10 обележја, узимају се сва обележја).

Затим је потребно испитати међусобну корелацију између изабраних обележја и навести која је метода коришћена. Приказати расподелу вредности корелационих коефицијената за изабрана обележја.

На крају, приказати график расподеле узорака по класама за два обележја која показују највећу корелацију са класом.

Редукција димензија [12]

Применити LDA (*Linear Discriminant Analysis*) на цео скуп обележја у циљу редукције димензија. На основу добијених резултата потребно је утврдити минималан број димензија на који оригинални скуп може бити редукован, тако да индекс информативности (односно кумулативни проценат објашњене варијансе) буде већи од 80%.

Добијене резултате је потребно прокоментарисати.

Тестирање хипотеза и параметарска класификација [15]

На скупу података након редукције димензија потребно је применити одговарајући класификациони приступ:

- у случају две класе – одабрати и применити параметарски класификатор у складу са степеном сепарабилности добијених података;
- у случају више класа – применити тест више хипотеза.

Уколико редукција димензија из претходног корака не обезбеди довољну сепарабилност класа, потребно је размотрити варијанту смањења броја обележја на основу њихове информативности, а затим пројектовати одговарајући класификатор заснован на тестирању хипотеза над таквим вишедимензионим подацима.

Резултате класификације потребно је приказати за независни тест скуп у облику конфузионе матрице, а затим на основу ње израчунати прецизност, сензитивност, специфичност и укупну тачност класификације.

Непараметарска класификација [20]

Потребно је изабрати једну непараметарску методу класификације и испитати њену перформансу на целокупном скупу података. Изабрани класификатор треба истренирати, а затим применом крос-валидације одредити оптималну вредност параметра:

- за KNN класификатор – одредити оптималан број суседа (K);
- за стабло одлучивања – одредити максималну дубину стабла.

приказати график зависности тачности класификације од оптимизованог параметра.

Резултате класификације потребно је приказати за независни тест скуп у облику конфузионе матрице, а затим на основу ње израчунати прецизност, сензитивност, специфичност и укупну тачност класификације.

Неуралне мреже [20]

На целокупном скупу података потребно је извршити обучавање и тестирање више структура неуралних мрежа, које илустровати по један пример добре структуре, структуре са премалим бројем неурона, и структуре са превеликим бројем неурона. Правилан избор структуре (оптималан број слојева и неурона) биће додатно бодован. За сваку структуру коју тренирате потребно је приложити график перформансе током обучавања.

У ситуацији када мрежа садржи превише неурона, неопходно је демонстрирати начине заштите од преобучавања. Потребно је укратко објаснити и применити следеће две мере:

- рано заустављање,
- регуларизација.

Резултате класификације потребно је приказати за независни тест скуп у облику конфузионе матрице, а затим на основу ње израчунати прецизност, сензитивност, специфичност и укупну тачност класификације.

Компаративна анализа [10]

У табеларној форми приказати резултате све три методе класификације које су примењене у оквиру претходних корака анализа (тест више хипотеза, параметарска, непараметарска и неуралне мрежа). За сваку од метода у табели навести прецизност, сензитивност, специфичност и укупну тачност класификације.

На основу добијених резултата, прокоментарисати предности и недостатке сваке од примењених метода и дати кратак закључак која од њих даје најбољу перформансу на анализираном скупу података.

Базе података

За сваки скуп података је дат .csv фајл у коме се налазе подаци. На датом линку се налази опис базе података као и потенцијално интересантне референце ка радовима. Ови радови могу да буду корисни са стране рангирања атрибута и додатних информација о бази.

Напомена: У случају тема 24-25 и 26-29 постоји више могућности избора излаза (назив за сваку тему је дефинисан у загради). Приликом класификације потребно је одбацити остале могућности излаза из скупа података.

1. Arrhythmia
   https://archive.ics.uci.edu/ml/datasets/Arrhythmia
2. Autistic Spectrum Disorder (Children)
   https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers
3. Bone Marrow Transplant
   https://archive.ics.uci.edu/ml/datasets/Bone+marrow+transplant%3A+children
4. Breast Cancer
   https://archive.ics.uci.edu/ml/datasets/Breast+Cancer
5. Breast Cancer Wisconsin
   https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
6. Mammographic
   https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass
7. Stroke Prediction
   https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset
8. Heart Disease
   https://archive.ics.uci.edu/ml/datasets/Heart+Disease
9. Heart Failure
   https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
10. Statlog Heart
    https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29
11. Hepatitis C
    https://www.kaggle.com/datasets/amritpal333/hepatitis-c-virus-blood-biomarkers
12. Contraceptive Method Choice
    https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice
13. Dermatology
    https://archive.ics.uci.edu/ml/datasets/Dermatology
14. Diabetic Retinopathy Debrecen
    https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set
15. Fertility
    https://archive.ics.uci.edu/ml/datasets/Fertility
16. HCV data
    https://archive.ics.uci.edu/ml/datasets/HCV+data

17. Thoracic Surgery
    https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data
18. Estimation of obesity
    https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+
19. Chronic Kidney Disease
    https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease
20. COVID presence
    https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence
21. Heart Disease Prediction
    https://www.kaggle.com/code/andls555/heart-disease-prediction
22. Cardiovascular Disease
    https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
23. Adult Autism
    https://www.kaggle.com/datasets/faizunnabi/autism-screening
24. Cardiotocography (излаз колона NSP)
    https://archive.ics.uci.edu/ml/datasets/Cardiotocography
25. Cardiotocography (излаз колона CLASS)
    https://archive.ics.uci.edu/ml/datasets/Cardiotocography
26. Cervical cancer (излаз колона Hinselmann)
    https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
27. Cervical cancer (излаз колона Schiller)
    https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
28. Cervical cancer (излаз колона Citology)
    https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
29. Cervical cancer (излаз колона Biopsy)
    https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
30. Hepatitis
    https://archive.ics.uci.edu/ml/datasets/Hepatitis
31. Stress
    https://www.kaggle.com/datasets/mdsultanulislamovi/student-stress-monitoring-datasets/data
32. Depression
    https://www.kaggle.com/datasets/adilshamim8/exploring-mental-health-data/data
33. Parkinsons
    https://archive.ics.uci.edu/dataset/174/parkinsons
34. Liver diseases
    https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset