Универзитет у Крагујевцу

Факултет инжењерских наука



**Системи за подршку одлучивању**

Пројектни задатак:

SPECTF Heart Data Set

Студент: Предметни наставник:

Тијана Танасковић 452/2021 Проф. Ненад Филиповић

Садржај:

[Увод 3](#_Toc95501689)

[Подаци 4](#_Toc95501690)

[Креирање модела подешавање хиперпараметара 6](#_Toc95501691)

[Закључак 11](#_Toc95501692)

[Литература 11](#_Toc95501693)

# Увод

Циљ овог пројектног задатка је креирање класификационог модела који ће на основу прикупљених података пацијенте да сврстава у две групе: нормалну и абнормалну.

Скуп података описује дијагностиковање слике срца техником једнофотонске емисионе компјутеризоване томографије (SPECT). Сваки пацијент је класификован у две категорије: нормалну и абнормалну.

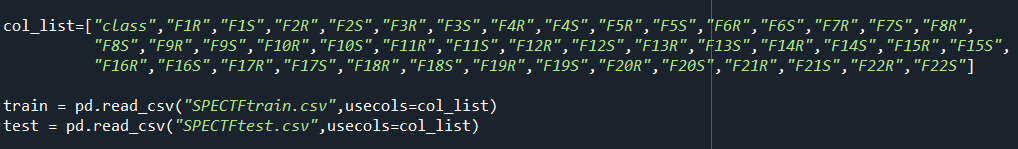
Скуп података има 44 улазна податка и један излазни (класу). Излазни може бити нула или јединица. Доступан је на [1].

# Подаци

Пре креирања класификационог модела, неопходно је прво извршити предпроцесирање и анализу скупа података.

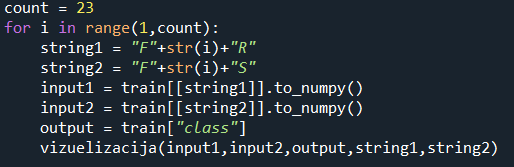
Предпроцесирање података је процес измене података, тако да сваки податак може да се искористи прибликом обучавања модела. Овај скуп података нема недостајуће вредности па није имало потребе вршити никакву измену.

Постоје два фајла од којих је први фајл са подацима које тренирано а у другом се налазе тест подаци. Потребно је прво њих учитати и пошто су већ подељени на тренинг и тест податке, ми не морамо да их делимо сами. На слици 1 је приказано учитавање података.

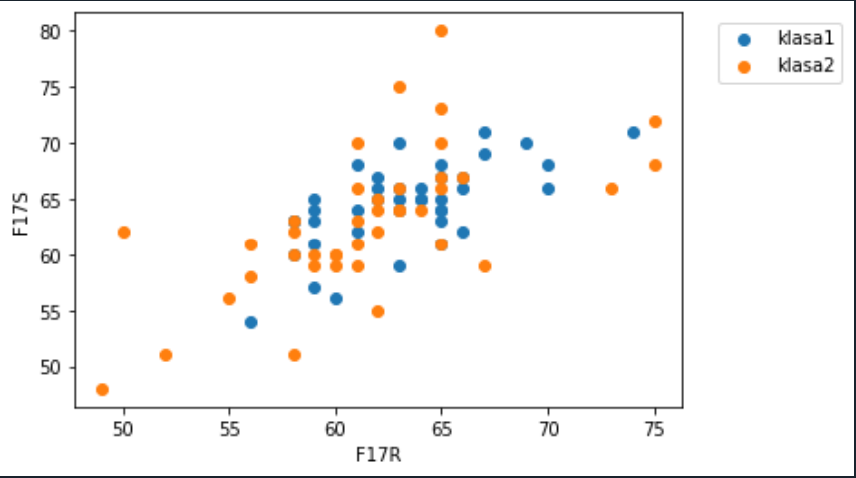
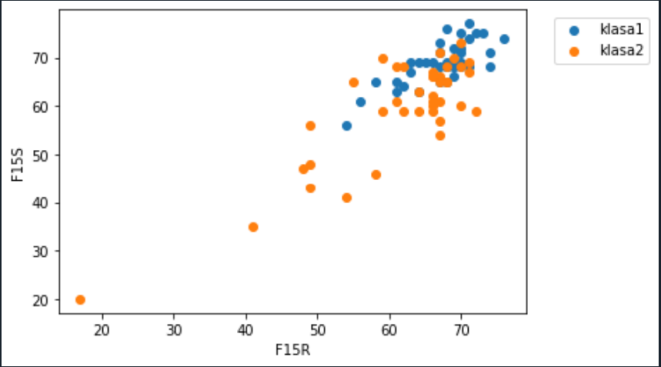
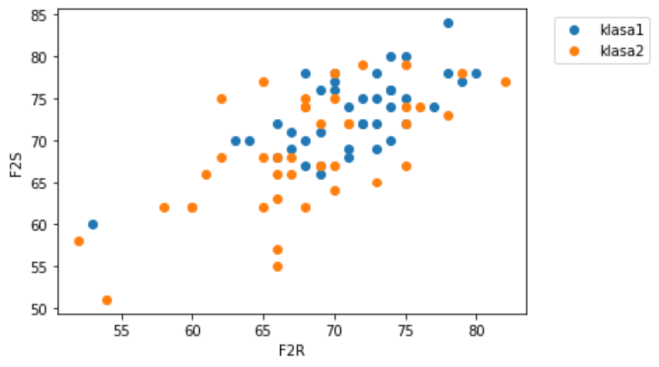


Слика 1. Учитавање података

Следећи корак је визуелизација података. Подаци су визуелизовани у паровима. На x оси налазе се подаци F1R, F2R, F3R ... F22R, док се на y оси налазе подаци F1S, F2S, F3S ... F22S. У коду је омогућено приказивање свих парова док ће само неки бити приказани даље у тексту.



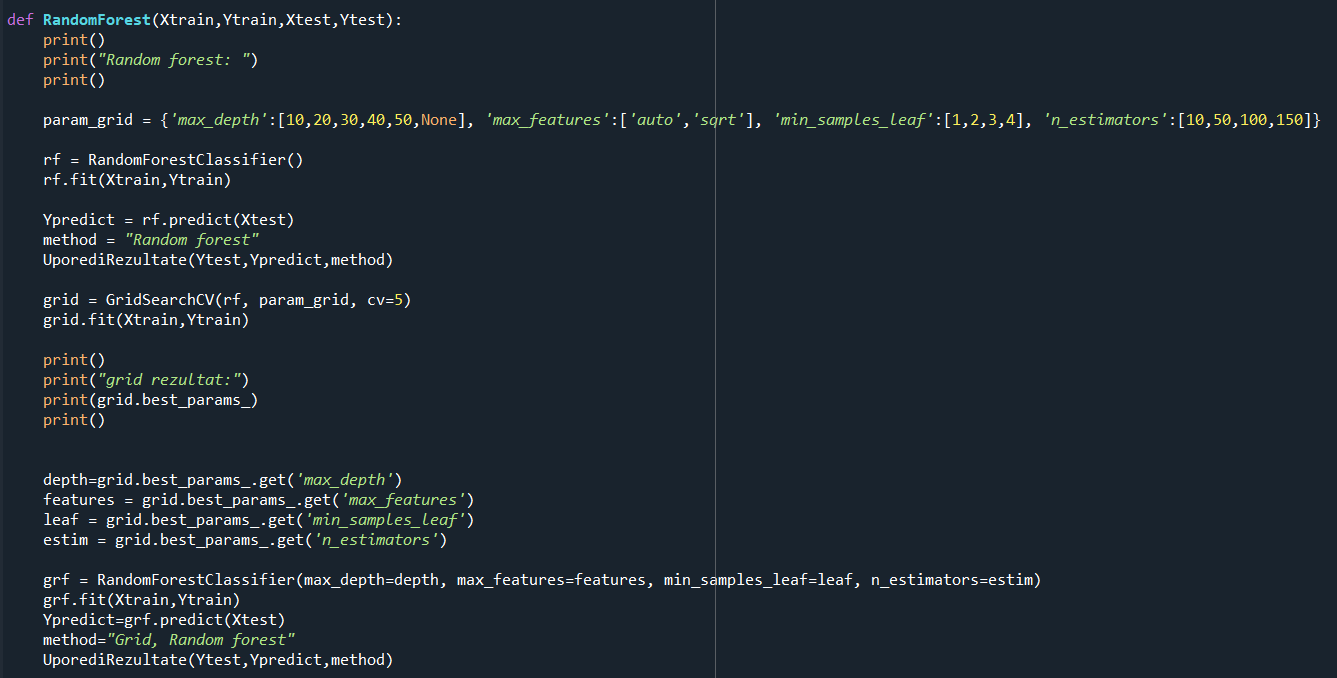
Слика 2. Визуелизација свих парова података



Слика 3. Визуелизација података

# Креирање модела подешавање хиперпараметара

Имплементирано је 6 алгоритама за решавање овог проблема и то су: логистчка регресија, Наивни Бајес, к најближих суседа, стабло одлучивања, random forest и support vector machine. ешавање проблема испробано је помоћу четири регресиона алгоритма: вишеваријантне линеарне регресије, полиномијалне линеарне регресије, стабла одлучивања и random forest регресора. У наставку рада биће речи о два алготирма који су дали највоље резултате и то су: random forest i support vector machine.



Слика 4. Дефинисање модела алгоритма Random Forest и подешавање хиперпараметара

Модел је дефинисан коришћењем методе RandomForestClassifier() библиотеке sklearn. Затим му се прослеђују тренинг скупови улазних и излазних података, да би се модел обучио. Онда се предвиђају излазне вредности за улазне вредности тест скупа и на самом крају се упоређују добијени резултати са жељеним резултатима.

Што се тиче подешавања хиперпараметара то је у ствари испробавање више различитих вредности хиперпараметара како би се дошло до закључка које хиперпараметре је најбоље користити.

Најважнији хиперпараметри овог алгоритма су: max depth, max features, min sample leaf и n estimators. Max depth је параметар који ограничава највећу могућу дубину стабла. Min sample leaf је параметар који дефинише минималан број одбирака потребан за креирање листа на дну стабла. Последице премалог и превеликог броја су исте. Max features ограничава величину стабла, чиме се подиже могућност боље генерализације, али се самим тим и повећава време потребно за извршавање алгоритма, што може представљати проблем када су у питању већи скупови података. N estimators параметар дефинише колико стабала одлучивања ће бити коришћено приликом обучавања модела. [2]

За подешавање хиперпараметара се користи метода GridSearchCV() библиотеке sklearn. Ова метода захтева прослеђивање имена параметара који ће бити испитани, као и предефинисаних вредности тих параметара које ће бити тестиране. Алгоритам затим креира све могуће комбинације вредности ових параметара и покушава да обучи модел, памтећи његове перформансе. Када се све комбинације испитају, ова метода ће вратити списак најбољих вредности, које се касније могу користити при креирању новог модела. Неопходно је вршити испитивање свих параметара истовремено због тога што они нису сви међусобно независни и најбоља могућа вредност се може разликовати у зависности од тога који параметри се још испитују и са којим вредностима.

Након извршеног тренирања и теститања, формира се конфузиона матрица на основу добијених и тражених вредности.

Са леве стране су представљене жељене вредности излаза, а са доње стране добијене реалне вредности. Матрица конфузије је битан показатељ прецизности, али и других метрика класификатора. Из матрице можемо да прочитамо вредноси TP(true positive), TN(true negative), FP(false positive) и FN(false negative).

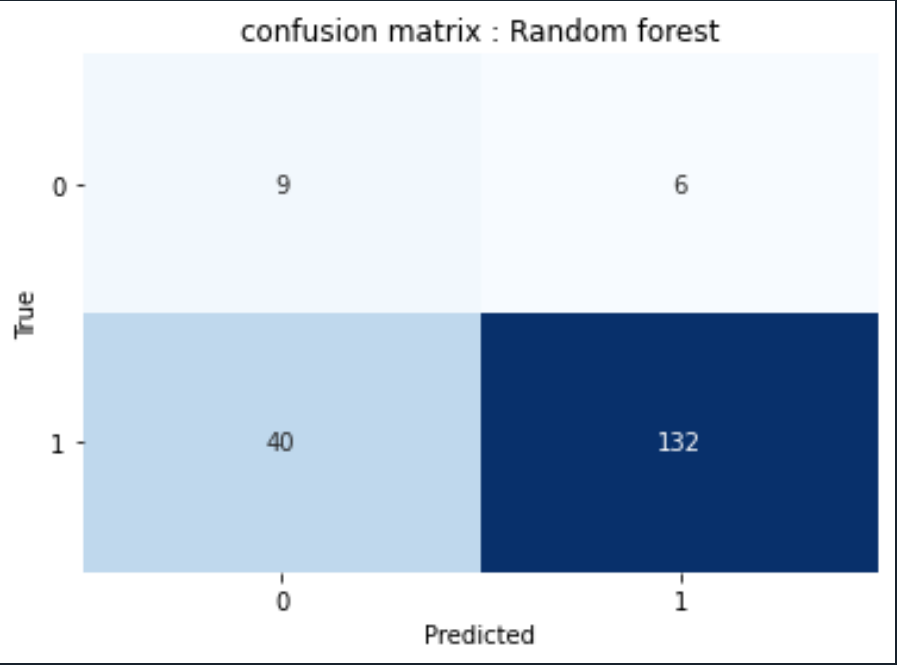
TP за класу 1 представља број пута који је мрежа на излазу добила класу 1, када је она била и жељени излаз.

TN за класу 1 представља број пута који је мрежа на излазу добила класу различиту од класе 1, када је и жељени излаз био различит од класе 1.

FP за класу 1 представља број пута који је мрежа на излазу добила класу 1, када она није била жељени излаз.

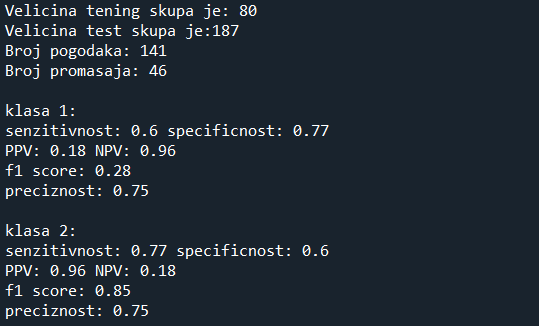
FN за класу 1 представља број пута које је мрежа на излазу добила класу различиту од класе 1, када је она била жељени излаз.

Помоћу ових вредности, даље, можемо да одредимо вредности метрика класификатора. Сензитивност представља способност класификатора да предвиђа TP за сваку класу, док специфичност представља способност класификатора да предвиђа TN за сваку класу. PPV(positive predictive value) и NPV(negative predictive value) за класу 1 представљају удео тачних резултата у укупном броју резултата када је на излазу тражена класа 1, односно када на излазу није тражена класа 1 респективно. Тачност је метрика која приказује удео TP и TN у укупном броју резултата, а f1-score додатна метрика прецизности која показује однос између TP и FP, као и однос између TP и FN.



Слика 5. Конфузиона матрица Random Forest алгоритма

Прецизност класификатора је 75%.



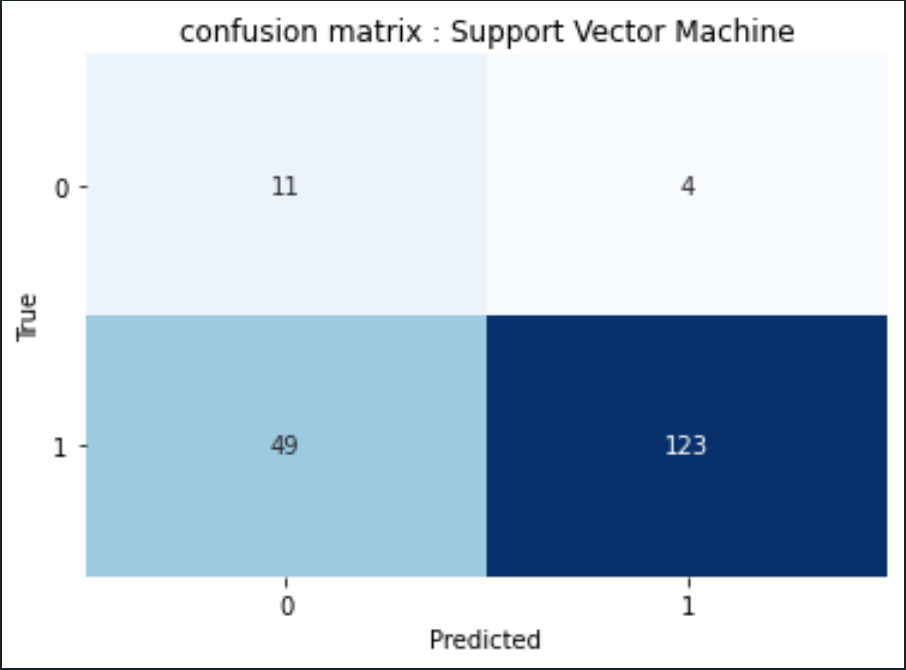
Слика 6. Добијене метрике коришћењем Random forest алгоритма

Support Vector Machine се имплементира на готово исти начин као и претходни алгоритам, само је битно напоменути који су то хиперпараметри који се користе код овог алгоритма.

Хиперпараметри овог модела јесу kernel функција и вредност C. Вредност С представља јачину регуларизације, односно у којој мери класификатор мора да се потруди да избегне било какве грешке приликом обучавања модела. Kernel функција дефинише облик равни којима се подаци раздвајају и може имати вредности linear и rbf. [3]

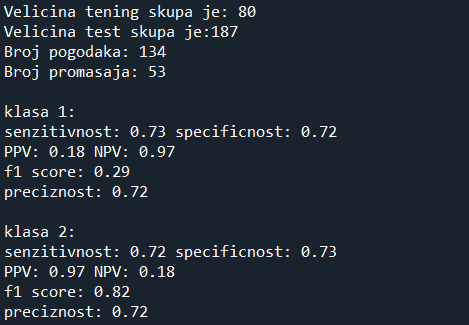


Слика 7. Дефинисање модела алгоритма Support Vector Machine и подешавање хиперпараметара



Слика 8. Конфузиона матрица Support Vector Machine алгоритма

Прецизност класификатора је 72%.



Слика 9. Добијене метрике коришћењем Support Vector Machine алгоритма

# Закључак

Након тестирања и подешавања хиперпараметара можемо закључити који алгоритми дају најбоље резултате. То су у овом случају random forest и support vector machine. Први са тачношћу од 75% а други са 72%. Бољу тачност алгоритама смо добили коришчењем grid searcha.

# Литература

[1] [UCI Machine Learning Repository: SPECTF Heart Data Set](http://archive.ics.uci.edu/ml/datasets/SPECTF+Heart) (10.02.2022.)

[2] [https://scikit-learn.org/stable/modules/generated/ sklearn.ensemble.RandomForestClassifier.html](https://scikit-learn.org/stable/modules/generated/%20sklearn.ensemble.RandomForestClassifier.html) : Random forest (10.02.2022.)

[3] <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html> : метода потпорних вектора (10.02.2022.)