PROIECT ISIA

Nume : Nicolae Alexandru Stefan

Setul de date repartizat constă în atribute sociodemografice dar si tipurile de asigurari pe care le detin instantele(tipurile de oameni, descrisi in functie de atributele mentionate anterior).

Setul de antrenament consta in 5822 de instante a cate 86 de atribute fiecare, cel de testare consta in 4000 de instante a cate 85 de atribute fiecare, lipsind ultimul atribut, cel ce trebuie prezis(CARAVAN : Number of mobile home policies).

Dupa analiza tipurilor de date, se va aplica SVR(Support Vector Regression).

Datorita numarului mare de atribute sociodemografice, se va renunta la o mare parte din aceasta deoarece nu ajuta la o clasificare eficienta.

In cazul implementarii acestui algoritm am folosit librariile: scikit-learn pentru implementarea SVR, numpy pentru utilizarea functiilor de incarcare a fisierelor text cat si a modificarii matricilor necesare, model\_selection derivata din libarira scikit-learn pentru a putea facea impartirea fisierelor in antrenare si testare si metrics pentru masurarea performantei. Deoarece este o problema de tip regresie, performanta nu se poate masura cu acuratatea, de aceea am folosit eroarea patratica medie(MSE) dar si ROC- pentru a observa procentajul de diferentiere intre predictiile calculate.

Mai jos se regasesc rezultate obtinute pentru variatiile costului:

|  |  |  |
| --- | --- | --- |
| COST | MSE | ROC |
| 0.03125 | 0.05758962463803557 | 52.95% |
| 0.125 | 0.05758158213207767 | 63.61% |
| 0.5 | 0.05757831707556596 | 66.06% |
| 2 | 0.057589066009414325 | 56.39% |
| 8 | 0.057574526495805925 | 62.14% |
| 32 | 0.05758103629886628 | 62.98% |
| 128 | 0.05759437817385931 | 48.92% |

Evidentiat se observa cea mai mica eroare, aceasta gasindu-se pentru cost = 8, si cea mai buna diferentiere a predictiilor pentru cost = 0.5.