Clasificare binară - AVC și Salarii

Cristian-Ștefan Avram
[1] *Facultatea de Automatică și Calculatoare - 331 CA*

26 May 2024

1 INTRODUCERE

Acest document suport pentru algoritmii de clasificare binară folosiți peste două seturi diferite de date, unul pentru AVC, altul pentru salarii, urmărește analiza amănunțită a rezultatelor obținute în urma antrenării unor modele de învățare automată. Deoarece sunt mai mulți pași implicați în procesul de antrenare și folosire a algoritmilor de clasificare, se vor explica, pe rând, pentru fiecare set de date, modificările aduse asupra datelor, pornind de la etapa de explorare a acestora, până la cea de preprocesare și, în sfârșit, cea de clasificare.

2 SALARII

2.1 Explorarea și analiza datelor

În etapa de explorare a datelor, foarte multe puncte cheie pot fi extrase din graficele si tendințele din setul de date cu salarii. Este foarte important contextul socio-economic atunci când se discută clasificarea unui angajat într-una din cele două categorii ($\leq 50K$ sau $> 50K$).

Pentru început, setul de date cuprinde foarte multe intrări pentru angajații ce au un salariu cel mult egal cu 50K, acestea fiind de aproape 4 ori mai multe decât cele reprezentative pentru cealaltă categorie, la fel cum se poate observa în Figura 1. De asemenea, se observă o distribuție normală între tipurile de atribute folosite pentru descrierea setului de date.

În rândul atributelor numerice, acestea urmează, de asemenea, o distribuție normală, fiind necesară eliminarea unor outliere provenite din erori de introducere a datelor, ori din existența foarte specifică a anumitor informații. Foarte important de menționat este faptul că majoritatea datelor sunt reprezentative pentru oameni cu vârste cuprinse între 30 și 50 de ani, iar numărul mediu de ore de lucru pe săptămână este aproximativ de 40 ore, după cum se poate observa și în Figura 3.

Pentru atributele categorice, se observă faptul că două treimi din date sunt relevante pentru sexul masculin (Figura 9), iar majoritatea datelor sunt pentru persoane etichetate prin ”albe” din punctul de vedere al rasei (Figura 7). De asemenea, majoritatea locurilor de muncă sunt din sectorul privat (Figura 5) și sunt în US (Figura 14). Pentru celelalte atribute, distribuțiile sunt așteptate, având în vedere, spre exemplu, tipul de educație, fiind de așteptat faptul că numărul de persoane cu un nivel mai mare de pregătire profesională și academică scade odată cu numărul de ani și cu experiența (Figura 13).

În matricea de corelație pentru atributele numerice din Figura 17, se observă o corelație foarte mare între atributele „prod” și „gain”. De asemenea, în matricea de corelație pentru atributele categorice (Figura 18), folosind metoda chi-pătrat, aproape toate atributele sunt corelate între ele.

2.2 Preprocesarea datelor

În etapa de preprocesare a datelor, este foarte importantă decizia de folosire a anumitor atribute din cele prezente în setul de date, prin intetmediul matricelor de corelație. Astfel, pentru o mai bună antrenare a modelului, am hotărât să renunț la următoarele atribute: „gain”, „partner”, „edu”, „race”, „gtype” și „work_type”.

Pe lângă decizia de eliminare a acestor variabile, am modelat și valorile celorlalte atribute din motive diverse. Pentru început, am eliminat outlierele existente, am standardizat valorile numerice si codificat atributele categorice.

Pentru atributul „country”, 89.8% din date erau reprezentative pentru „US”, după cum se poate observa în Figura 16. Având în vedere acest lucru, am decis să modific valoarea intrărilor care nu erau „US” și să le clasific în „Non-US” pentru departajarea țărilor – US și țări care nu sunt US (știind că există 40 de țări unice). Decizia respectivă a fost luată și din pricina faptului că se poate observa că în țările din afara US salariul este mai mic (Figura 15).

De asemenea, am împărțit valorile atributului „work_type” în „gov” și „self”, reprezentând dacă job-ul este guvernamental sau nu, după observarea distribuției.

2.3 Evaluarea algoritmilor de clasificare

2.3.1 Regresie Logistică

Pentru algoritmul de regresie logistică am obținut următoarele performanțe:

- Regresie logistică implementată manual (Figura 19)
  - Acuratețe - train: 0.813, test: 0.817
  - Precizie - 0.66
  - Recall - 0.47
  - F1 - 0.551
- Regresie logistică Sklearn (Figura 20)
  - Acuratețe - train: 0.83, test: 0.826
  - Precizie - 0.69
  - Recall - 0.53
  - F1 - 0.59

Se observă faptul că rezultatele sunt similare. De asemenea, pentru algoritmul implementat manual se pornește mereu cu ponderi aleatorii, iar rezultatele prezentate sunt de fapt o medie. Pentru clasificatorul Sklearn a fost folosit un „random state” pentru a obține aceleași rezultate. Cred că algoritmul Sklearn este puțin mai performant datorită găsirii unei frecvențe de învățare („learning rate”) mai bune.

2.3.2 *MLP*

Pentru MLP am obținut următoarele performanțe:

- MLP - manual (Figura 21)
  - Acuratețe - 0.837
  - Precizie - 0.72
  - Recall - 0.53
  - F1 - 0.61
- MLP - Sklearn (Figura 22)
  - Acuratețe - 0.82
  - Precizie - 0.73
  - Recall - 0.72
  - F1 - 0.72

La fel ca în cazul regresiei liniare, acuratețea obținută de ambii algoritmi sunt similari. Se observă faptul că acuratețea obținută în cadrul agloritmului implementat manual este puțin mai bună. De asemenea, în Figura 23 se pot observa și curbele de eroare și performanță, de unde se poate concluziona faptul că modelul nu face overfitting.

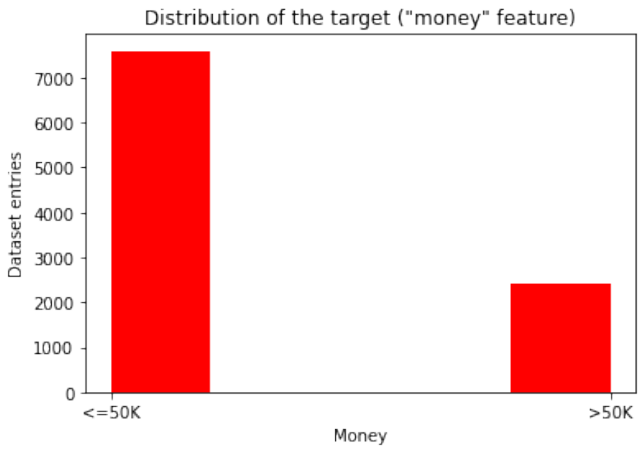

Figure 1.

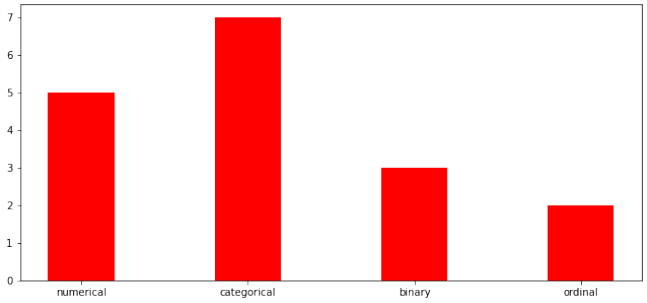

Figure 2.

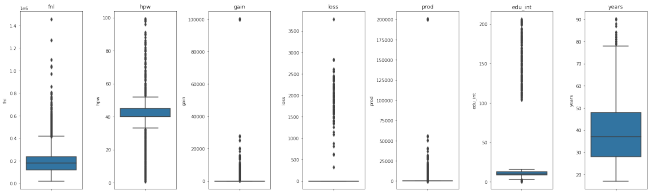

Figure 3.

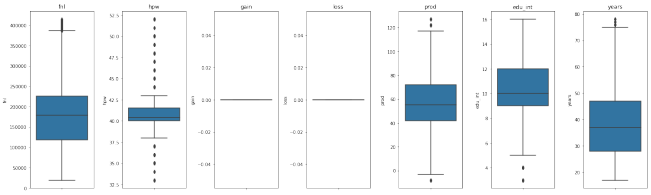

Figure 4.

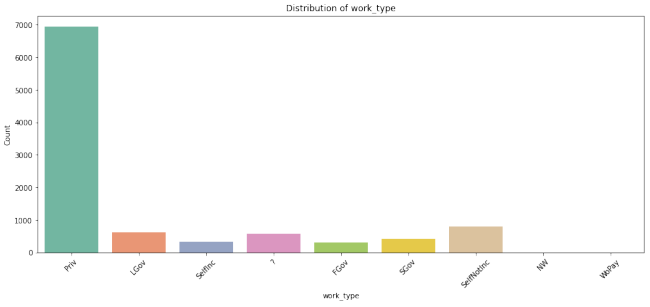

Figure 5.

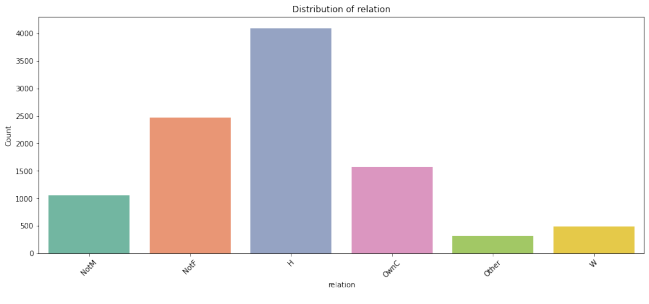

Figure 6.

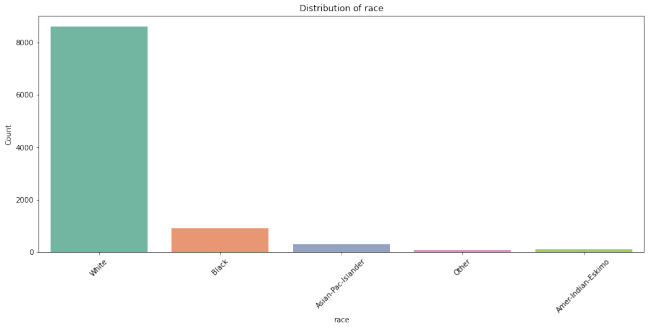

Figure 7.

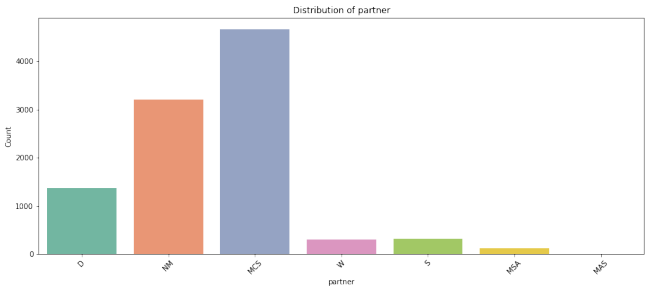

Figure 8.

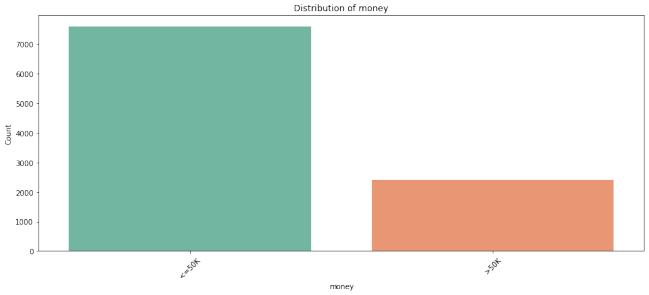

Figure 9.

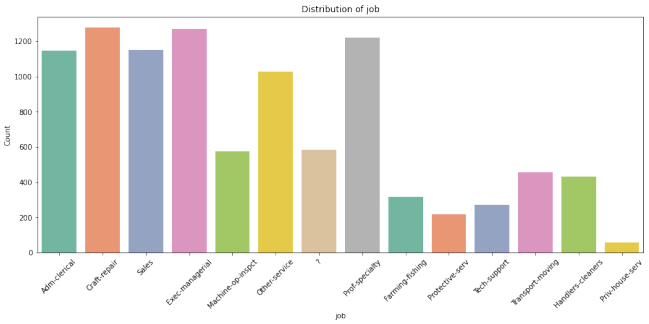

Figure 10.

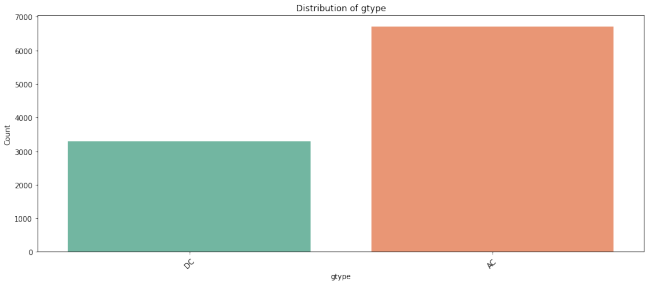

Figure 11.

Figure 12.

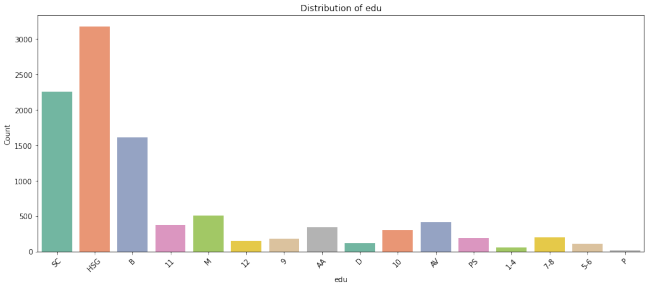

Figure 13.

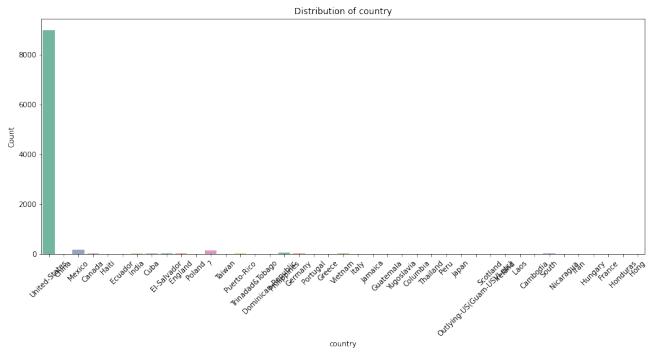

Figure 14.

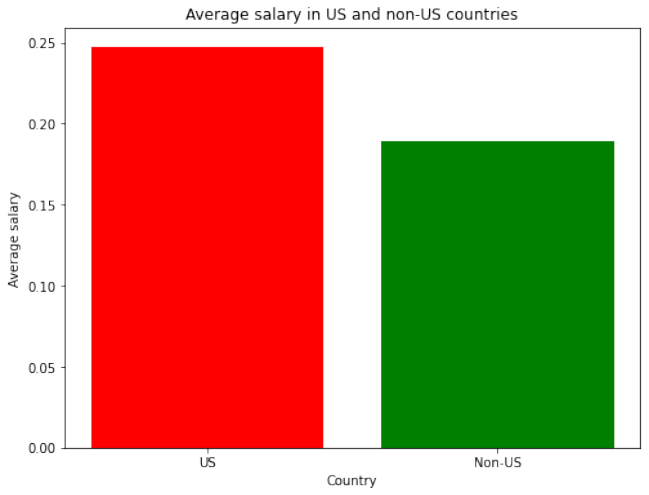

Figure 15.

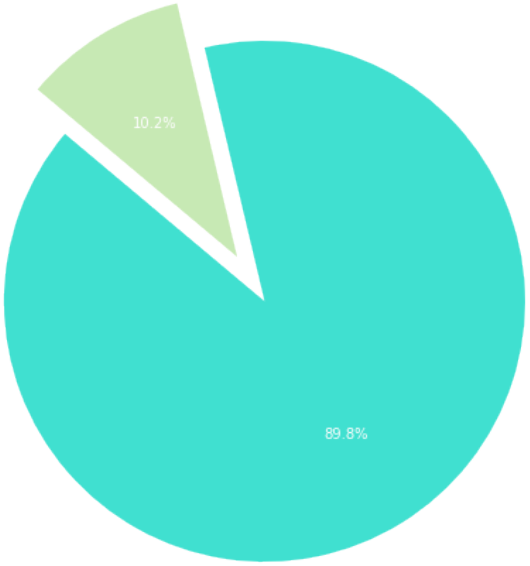

Figure 16.

Figure 17.

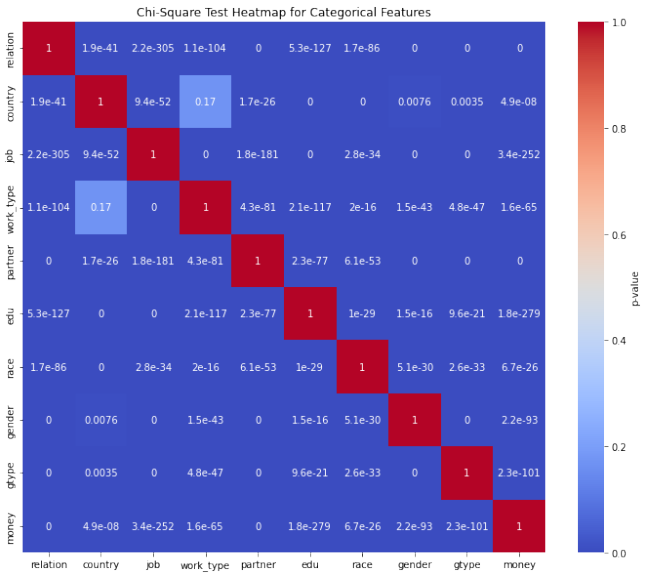

Figure 18.

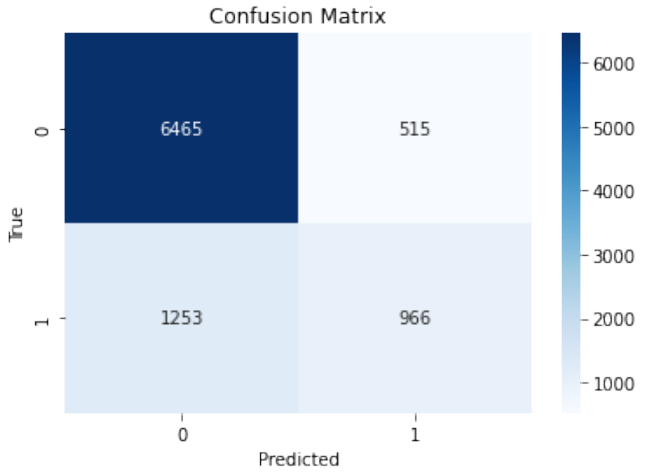

Figure 19. Salarii Regresie logistică - manual

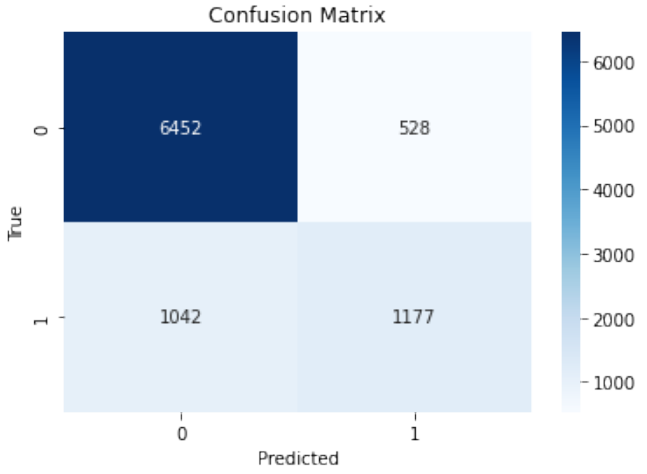

Figure 20. Salarii Regresie logistică - Sklearn

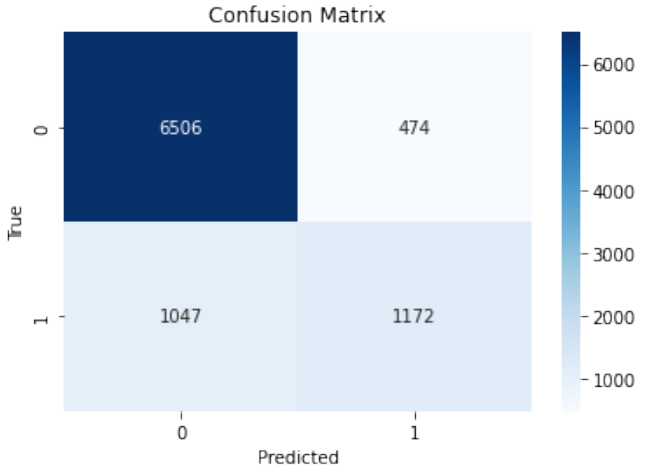

Figure 21. Salarii MLP manual

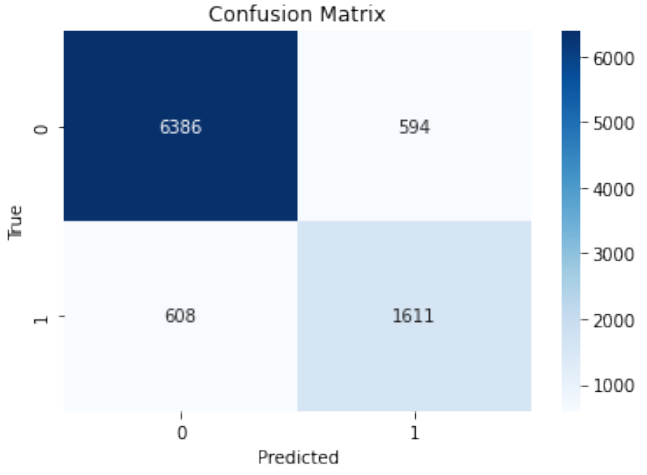

Figure 22. Salarii MLP Sklearn

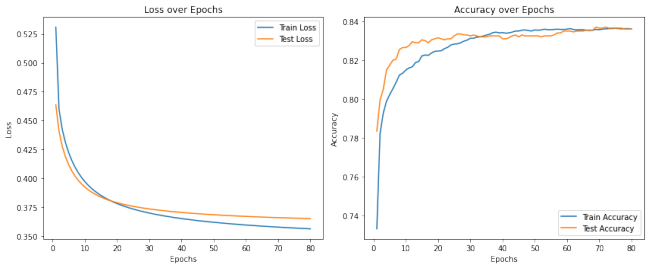

Figure 23. Salarii MLP - curbe de eroare și performanță

3 AVC

3.1 Explorarea și analiza datelor

Încă de la început, se poate observa faptul că pentru persoanele care au suferit AVC există mult prea puține intrări în setul de date (Figura 24). Acest aspect al setului de date sugerează riscul de overfitting în antrenarea modelului, întrucât doar 4.9% din oamenii prezenți în setul de date au suferit AVC (Figura 25).

În continuare, foarte multe atribute sunt binare, însemnând că mulțimea lor de valori are un cardinal egal cu 2. Acest fapt facilitează codificarea și folosirea acestor resurse prin clasarea valorilor cu 0 și 1.

În Figura 27 se poate observa distribuția valorilor pentru atributele numerice. Este important de notat faptul că majoritatea persoanelor din setul de date au vârste cuprinse între 20 și 60 de ani. De asemenea, sunt destul de puține persoane care au BMI mare, însemând ca persoanele din setul de date sunt preponderent sănătoase din punctul de vedere al greutății corporale, adică obezitatea nu ar fi neapărat un factor foarte important pentru acest set de date.

În Figura 29 sunt reprezentate datele categorice. Este foarte important faptul că majoritatea persoanelor sunt din sectorul privat, unde se preconizează că există această înclinație către fumat „social". Despre distribuția de fumători, aproximativ jumătate dintre persoane au fumat la un moment dat. Acest aspect este foarte important pentru clasificator, întrucât, după cum se poate observa și în figurile 30 și 31, 45% dintre persoanele care au suferit de AVC au fumat la un moment dat.

Se poate observa, de asemenea, și faptul că procentele de oameni care au probleme cardiovasculare, au tensiunea arteriala mare și un somn neregulat sunt foarte asemănătoare. Distribuțiile categorice sunt suficient de normale.

În matricea de corelație a atributelor numerice (Figura 32), se pot observa mai multe puncte cheie. În primul rând, somnul neregulat este asociat cu un risc crescut de probleme cardiovasculare, ceea ce este normal, întrucât somnul este foarte important pentru sănătatea inimii. De asemenea, atributul pentru rezultatele analizelor este foarte corelat cu cel pentru nivelul de zahăr din sânge, ceea ce este evident, având în vedere că analizele pentru sistemul cardiovascular includ și testarea nivelului de zahăr din sânge (riscul de dezvoltare a altor boli – ex: diabet). Alt punct cheie este reprezentat de faptul că vârsta biologică este foarte corelată cu vârsta calendaristică a unei persoane, ceea ce este evident. Cu toate acestea, vârsta biologică este mult mai importantă, întrucât reprezintă de fapt cât de sănătos este individul. De asemenea, sunt câteva corelații între vârstă și starea de relație a persoanelor, întrucât oamenii tind să se căsătorească la vârste mai înaintate.

Pentru atributele categorice (Figura 33), se poate observa doar o corelație foarte mare între job-ul unei persoane și dacă aceasta fumează. Se poate spune că unele locuri de muncă implică mai mult stres, ce este foarte legat în „popor" de consumul de tutun. Cu toate acestea, acest lucru poate desemna doar o simplă coincidență în setul de date.

3.2 Preprocesarea datelor

Pentru etapa de preprocesare a datelor am eliminat outlierele sau le-am modificat. De asemenea, m-am folosit de un „StandardScaler()" pentru standardizarea datelor. Codificarea am realizat-o folosind OHE și Label Encoding.

Atributele eliminate sunt: „mean blood sugar level", „years old", „chaotic sleep".

3.3 Evaluarea algoritmilor de clasificare

3.3.1 Regresie Logistică

Pentru algoritmul de regresie logistică am obținut următoarele performanțe:

- Regresie logistică implementată manual (Figura 19)
  - Acuratețe - train: 0.955, test: 0.925
  - Precizie - 0.083
  - Recall - 0.004
  - F1 - 0.007
- Regresie logistică Sklearn (Figura 20)
  - Acuratețe - train: 0.956, test: 0.925
  - Precizie - 0.0
  - Recall - 0.0
  - F1 - nan

Se observă faptul că acuratețea obținută este una foarte mare, iar performanțele algoritmilor sunt foarte asemănătoare.

3.3.2 MLP

Pentru MLP am obținut următoarele performanțe:

- MLP - manual (Figura 21)
  - Acuratețe - 0.95
  - Precizie - nan
  - Recall - 0.0
  - F1 - nan
- MLP - Sklearn (Figura 22)
  - Acuratețe - 0.93
  - Precizie - nan
  - Recall - 0.0
  - F1 - nan

Se observă faptul că acuratețea obținută este una foarte mare, iar performanțele algoritmilor sunt foarte asemănătoare.

3.3.3 Motivație

Deși acuratețea este una foarte mare pentru toți algoritmii, valorile preciziei, recall și F1 evidențiază aspectul menționat la început, faptul că modelul face overfitting. Acest lucru se întâmplă deoarece sunt mult prea puține informații despre persoanele care suferă de AVC, iar modelul tinde să aibă un bias față de de persoanele care nu au riscul de a suferi de AVC. De aceea, acuratețea este foarte mare, întrucât modelul prezice majoritatea cazurilor bine, dar nu este capabil să surprindă faptul că un om poate suferi sau nu de AVC.

O soluție pentru această problemă ar fi introducerea mai multor date pentru persoanele care au suferit de AVC ori reducerea numărului de intrări ale persoanelor care nu au suferit de AVC. Ultima abordare, care ar putea fi realizată ușor, aduce cu sine o altă problemă, ci anume o insuficiență de date pentru claritatea modelului în clasificarea posibilității ca o persoană să sufere de AVC.

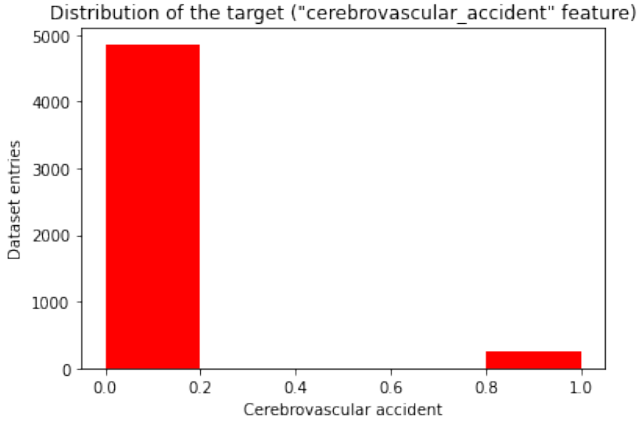

Figure 24.

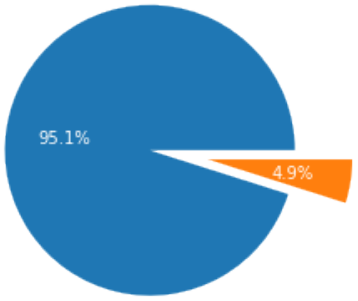

Figure 25.

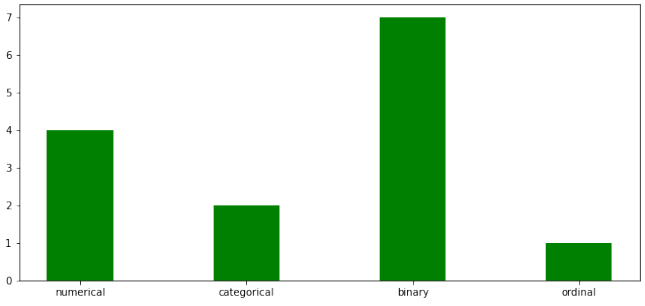

Figure 26.

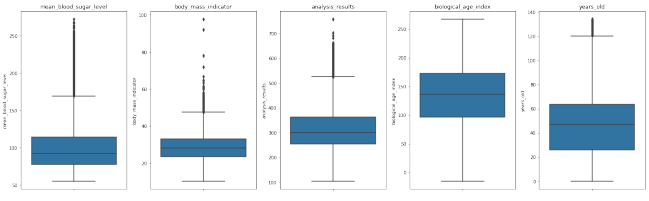

Figure 27.

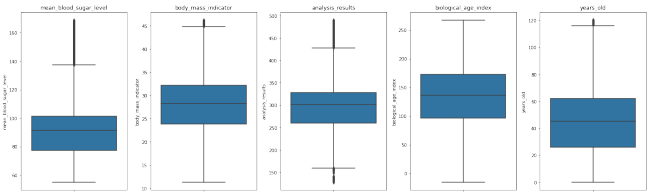

Figure 28.

Figure 29.

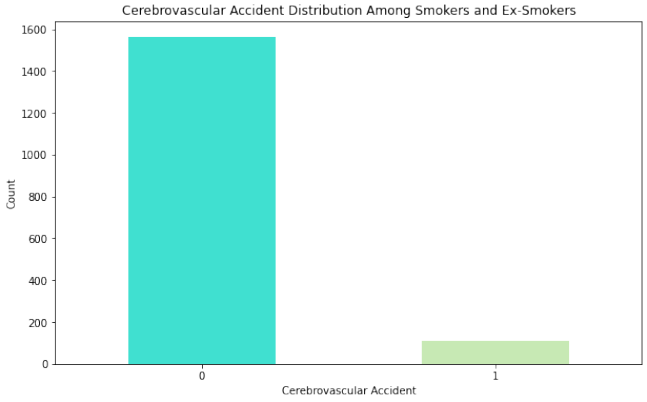

Figure 30.

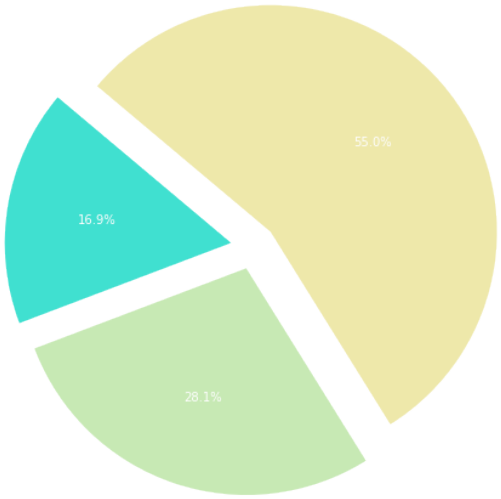

Figure 31.

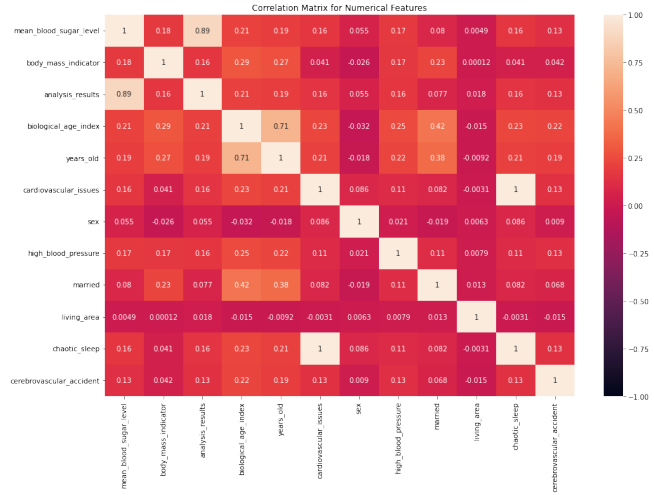

Figure 32.

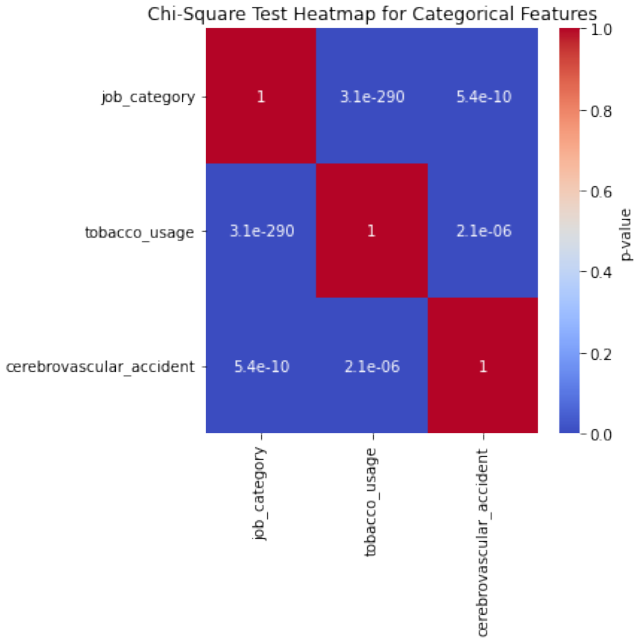

Figure 33.

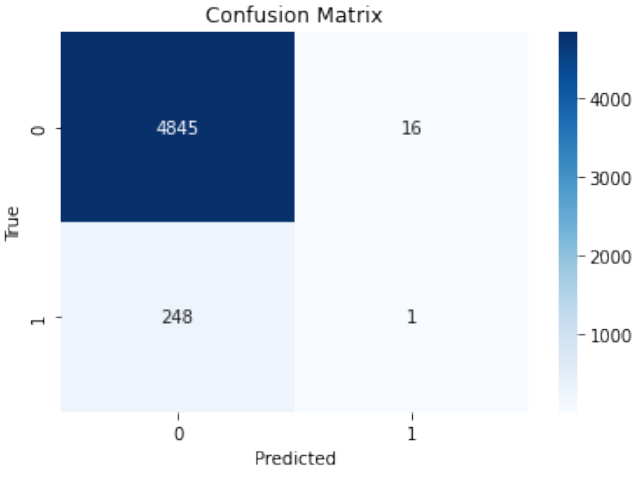

Figure 34. AVC Regresie Logistică manual

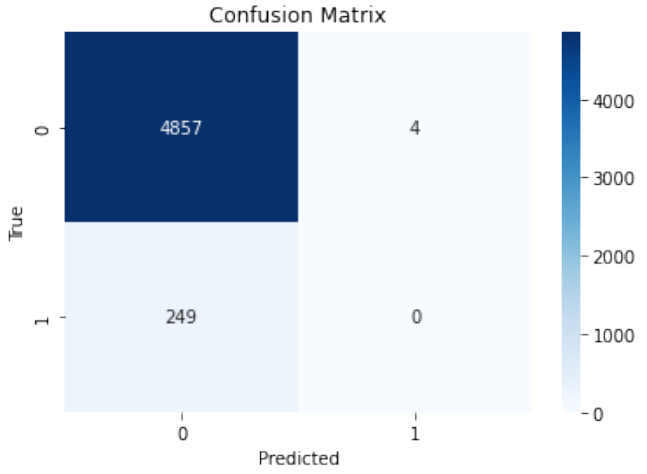

Figure 35. AVC Regresie Logistică Sklearn

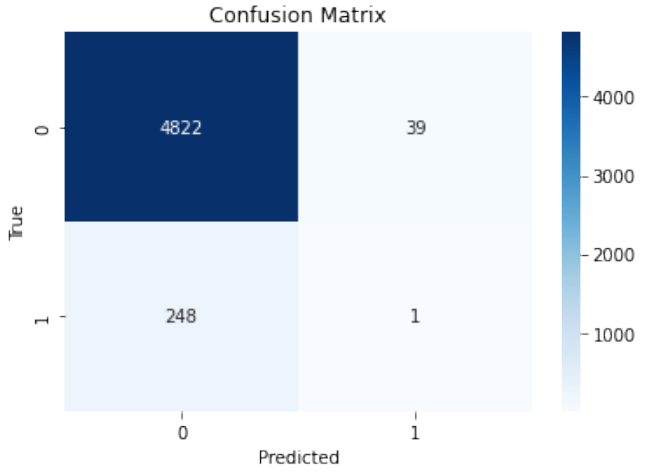

Figure 36. AVC MLP manual

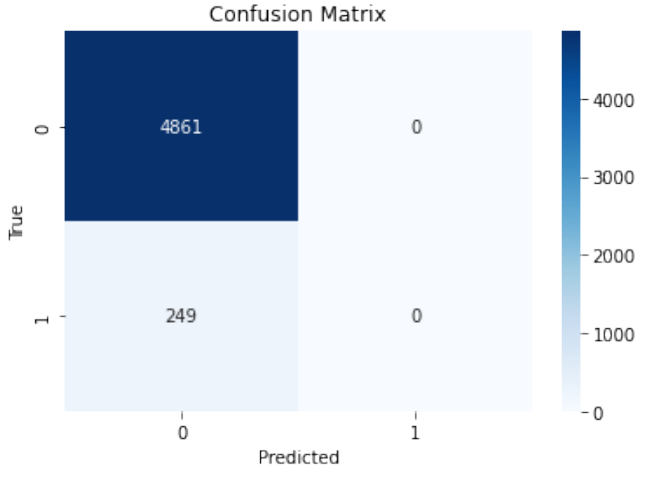

Figure 37. AVC MLP Sklearn

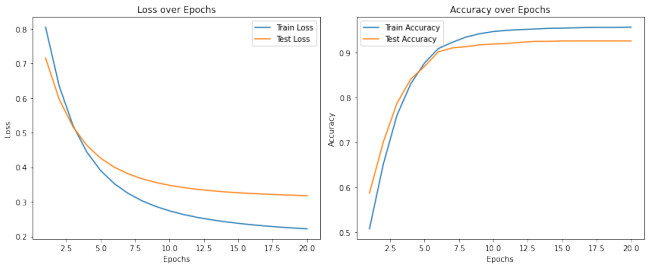

Figure 38. AVC MLP - curbe de eroare și performanță

This paper has been typeset from a TEX/LATEX file prepared by the author.