Clasificarea consumului de drogurilor

Dorin Zaharie

May 2024

1 Introducere

Consumul de droguri este o problemă globală cu implicații semnificative asupra sănătății publice, siguranței sociale și economiei. În ciuda eforturilor ample de prevenție și intervenție, consumul de droguri rămâne o provocare majoră, afectând negativ viețile a milioane de oameni din întreaga lume. De la substanțe ilegale precum heroina și cocaina, până la substanțe legale cum ar fi alcoolul și tutunul, consumul de droguri poate avea consecințe devastatoare pentru oameni.

În acest context, analiza datelor despre consumul de droguri devine crucială pentru înțelegerea și abordarea acestei probleme. Bazele de date referitoare la consumul de droguri oferă o sursă importantă de informații despre modelele de consum, factorii de risc și impactul asupra sănătății și societății. Prin analiza acestor date, se poate ajunge la soluții productive pentru a diminua aceste vicii.

Proiectul se concentrează pe explorarea și analiza unei astfel de baze de date. Această bază de date cuprinde informații detaliate despre consumul de droguri în rândul unei populații diverse, incluzând variabile legate de demografie, personalitate și modele specifice de consum de droguri. Scopul este să investigăm aceste date pentru a identifica factorii care influențează deciziile legate de consumul de droguri și pentru a dezvolta modele predictive eficiente.

Pornind de la această analiză, ne propunem să răspundem la întrebările tot mai dese care apar in privința consumului de droguri. Prin intermediul acestui proiect, sperăm să aducem o contribuție semnificativă la înțelegerea și abordarea consumului de droguri, contribuind astfel la promovarea sănătății și bunăstării societății în ansamblu.

2 Context

2.1 Descrierea bazei de date

Baza de date pe care am folosit-o conține informații relevante și detaliate despre consumul de droguri în rândul unei populații diversificate. Prin agregarea acestor date, baza de date oferă o imagine complexă și actualizată a comportamentului legat de consumul de droguri în diverse contexte și grupuri de populație. Baza de date conține o gamă variată de variabile, inclusiv demografice (cum ar fi

vârsta, genul, educația), de personalitate (precum scorurile Nscore, Escore, Oscore etc.) și modele specifice de consum de droguri (cum ar fi frecvența consumului de cannabis, cocaină etc.). Link-ul bazei de date:https://www.kaggle.com/datasets/mexwell/drug-consumption-classification/data

2.2 Obiective

Proiectul își propune să utilizeze această bază de date pentru a investiga și a analiza comportamentul legat de consumul de droguri și factorii care îl influențează. Dorim să obținem informații despre modelele de consum de droguri, factorii de risc și impactul asupra sănătății și societății. Prin intermediul analizei datelor, ne propunem să identificăm corelații și modele predictive care să ne ajute să înțelegem mai bine fenomenul consumului de droguri și să dezvoltăm strategii eficiente de prevenire acestuia. Prin intermediul acestui proiect vrem să răspundem la întrebări precum: Care sunt factorii care contribuie la decizia unei persoane de a consuma droguri? Cum pot fi identificate și evaluate modelele și corelațiile dintre caracteristicile individuale și comportamentul de consum de droguri? Cum putem utiliza aceste informații pentru a dezvolta strategii și politici eficiente de prevenire și intervenție în consumul de droguri?

2.3 Analiza

În baza de date pe care am prelucrat-o se regăsesc date despre 1888 de persoane, fiecare având diverse valori reprezentative. Distribuția genului din baza de date este prezentata în figura 1.

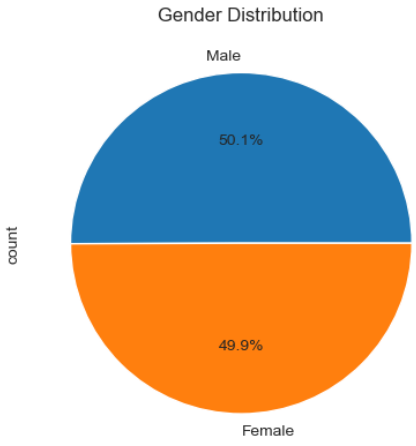

Figure 1: Distribuția genului

Distribuția vârstei este prezentată in figura 2. Distribuția educației pentru persoanele din baza de date este prezentată în figura 3. Fiecare valoare din

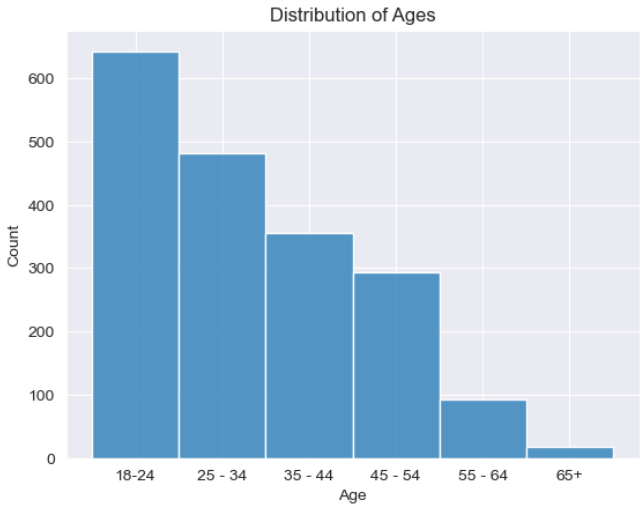

Figure 2: Distribuția vârstei

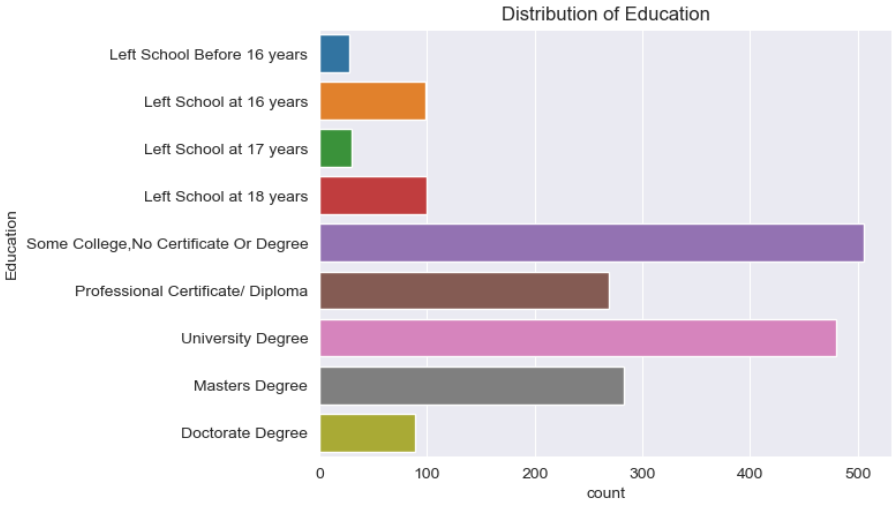

Figure 3: Distribuția educației

această bază de date o să ne ajute la clasificarea diferitor tipuri de droguri și la identificarea factorilor care influențează o persoană să le consume.

3 Aspecte teoretice relevante

3.1 Gini index

Gini index este o măsură a inegalității sau repartizării unei distribuții după cum spun Gastwirth și Joseph L [3] în articolul lor. În cadrul proiectului am calculat gini index-ul pentru aspectele demografice si scorurile de personalitate, fapt ce se poate observa în figura 4. Din acest calcul reiese datele sunt distribuite destul de uniform.

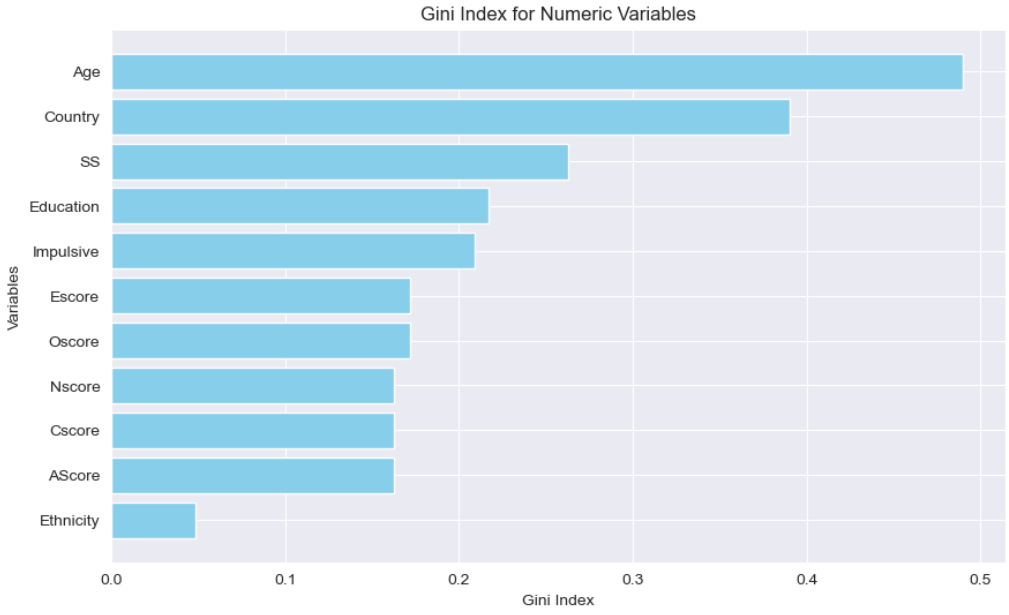

Figure 4: Gini index

3.2 Corelații

Există mai multe tipuri de corelații, printre cele mai comune fiind corelația Pearson, Spearman și corelația Kendall.Valoarea corelației este întotdeauna între -1 și 1. Cu cât valoarea este mai aproape de 1 respectiv -1, cu atât mai puternică este corelația. O corelație de 0 indică lipsa unei corelații liniare între variabile. În matricea prezentată în figura 5 se pot observa toate corelațiile existente in baza de date. Chiar dacă numărul corelațiilor este mare, sunt foarte puține cele care influențează datele.

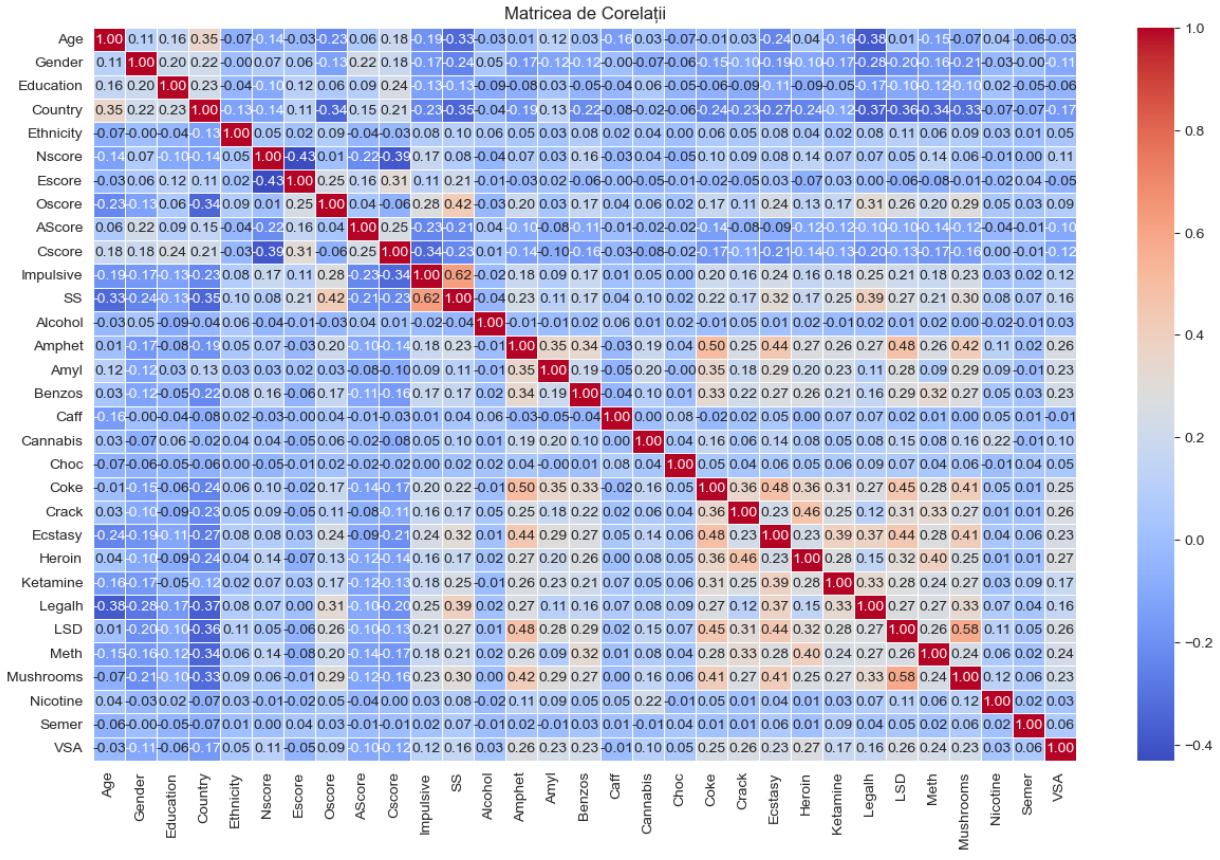

Figure 5: Matricea de corelații

Mark A. Hall [4] spune în lucrarea lui că o caracteristică ar fi utilă doar dacă este corelată cu sau predictivă a clasei, practic acesta insumează faptul că orice obiect sau eveniment poate fi utilizat pentru a-l descrie sau clasifica. Ca să fie corelație ar trebui să fie o asociere statistică semnificativă între o caracteristică și o clasă, iar ca sa fie predictivă, ar trebui să aibă abilitatea de a prezice cu exactitate la ce clasă aparține un obiect sau eveniment.

Corelațiile mai au o caracteristică importantă, aceea fiind simetria.

3.3 Entropia

Entropia reprezinta o masura a dezordonarii sau incertitudinii intr-un set de date. Cu cat probabilitatile sunt mai uniform distribuite, cu atat entropia este mai mare, in schimb daca aceasta este cat mai mica atunci setul de date este bine organizat. Originea termenului de entropie în acest domeniu a fost dată de Claude Elwood Shannon in lucrarea [10] sa. Din figura 6 reiese că baza de date este organizată bine.

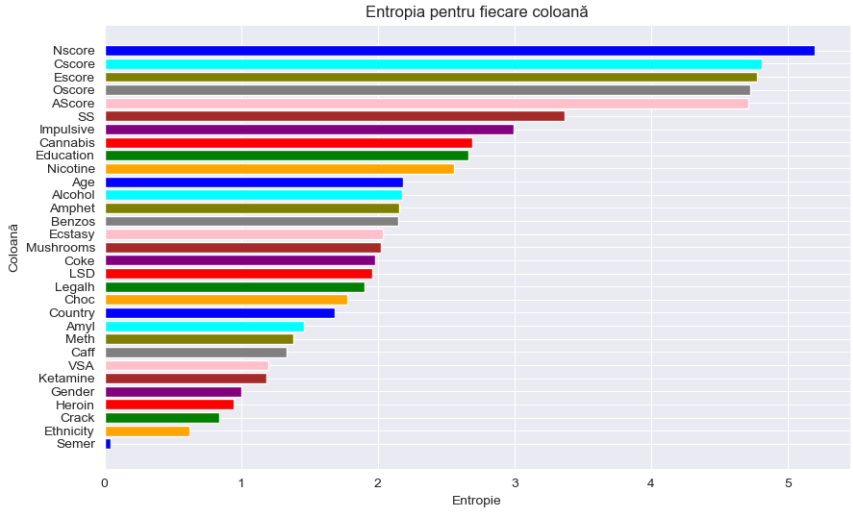

Figure 6: Entropia

3.4 Starea actuală a domeniului

Analiza consumului de droguri este un domeniu interdisciplinar complex care implică concepte și tehnici dintr-o varietate de domenii, inclusiv epidemiologie, psihologie, sociologie, statistică și informatică. Pentru a înțelege și a explora în profunzime comportamentul legat de consumul de droguri, este important să ne bazăm pe o înțelegere solidă a aspectelor teoretice și a cercetărilor anterioare din domeniu.

Howard Parker spune în articolul său [8] că deși ratele de consum de alcool per capita în Marea Britanie sunt departe de a fi cele mai mari la nivel internațional, Regatul Unit are cea mai implicată populație în consumul de droguri dintre toate statele Uniunii Europene. Acest statut nu este generat în primul rând de consumul problematic, deși dependența de heroină, crack și polidroguri se află la rate semnificative, ci de consumul de droguri "recreaționale" al tinerilor britanici (cu vârste între 12 și 30 de ani). Este vorba despre consumul de canabis, urmat de amfetamine, LSD, ecstasy și, din ce în ce mai mult, cocaină, care generează poziția de top a Marii Britanii în Europa atât per capita, cât și în mod specific printre populațiile tinere comparative.

Nils Braakmann și Simon Jones spun în articolul lor [2] că efectele consumului de canabis asupra unei populații vaste și reglementarea accesului la acesta, au fost discutate intens atât în public, cât și în mediul academic. Dovezile din diverse țări indică faptul că consumul de canabis poate duce la probleme de sănătate fizică și mentală, deși efectele nu sunt adesea mari; are efecte mixte asupra salariilor; are efecte negative, dar adesea slabe asupra (ne-)angajării și este corelat cu utilizarea altor droguri mai puternice, deși nu este clar în ce măsură canabisul cauzează consumul acestor droguri, așa cum este prezis de teoria gateway.

Consumul unei varietăți de substanțe naturale sau sintetice poate duce la dependență, fapt menționat de Christian Lüscher și Mark A Ungless în articolul [7] lor.Tot aceștia spun ca deși unele tipuri de droguri țintesc celule diferite din creier, toate au un efect inițial comun, acela de creștere a dopaminei.

În timpul adolescenței, interacțiunile cu colegii influențează atitudinile și comportamentele unui adolescent. Adolescenții caută aprobarea și acceptarea colegilor, ceea ce îi poate determina să se angajeze în comportamente riscante pentru sănătate, cum ar fi fumatul și consumul de droguri, lucru amintit de Lopez-Mayan et al [6].

Consumul de droguri este strâns legat de diverse probleme de sănătate mintală, inclusiv depresie, anxietate și tulburări psihotice. Este important de menționat că relația cauzală poate fi bidirecțională. Pe lângă efectele imediate, consumul de droguri poate avea consecințe pe termen lung asupra sănătății fizice și cognitive. Acestea includ deteriorarea funcțiilor cognitive, probleme cardiovasculare și riscul crescut de boli cronice.

4 Implementarea aspectelor teoretice

Pentru a atinge obiectivele propuse, am utilizat o serie de metode și tehnici din domeniul analizei datelor și al inteligenței artificiale, precum și pe o înțelegere solidă a cercetărilor anterioare în domeniu.

4.1 Prerlucrarea datelor

Pentru început, am prelucrat și curățat datele din baza de date "DrugConsumptionQuantified", asigurându-ne că acestea sunt valide și pot fi utilizate în analiza

noastră. Acest proces de prelucrare a datelor a inclus eliminarea datelor lipsă, standardizarea și normalizarea datelor și transformarea variabilelor categorice în variabile numerice, acolo unde a fost necesar.

Am aplicat diferite tehnici de analiză pentru a investiga relațiile între variabilele din setul nostru de date. Am utilizat modele predictive și clasificatoare pentru a anticipa comportamentul de consum de droguri pe baza caracteristicilor individuale și de personalitate.

Modelele de machine learning pe care le-am utilizat în cadrul acestui proiect au fost Logistic Regression, Random Forest și XGBoost.

4.2 Tehnologii

Pentru realizarea acestui proiect am folosit DataSpell ca și mediu de dezvoltare, Jupyter Notebook pentru scrierea codului si vizualizarea rezultatelor și Phyton3 ca și interpretor.

4.3 Modele

Regresia logistică este un algoritm de învățare automată supravegheat utilizat pentru sarcini de clasificare unde scopul este de a prezice probabilitatea ca o instanță să aparțină unei anumite clase sau nu. Regresia logistică este un algoritm statistic care analizează relația dintre doi factori de date [5].

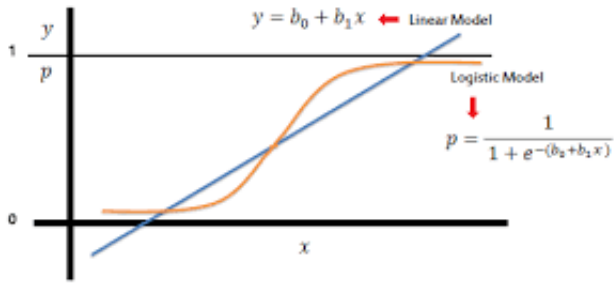

Figure 7: Logistic regression

Random forest este un algoritm de învățare supervizată. "Pădurea" pe care o construiește este un ansamblu de arbori de decizie, de obicei antrenați cu metoda de îmbinare. Ideea generală a metodei de îmbinare constă în faptul că o combinație de modele de învățare crește rezultatul general [9]. Comportamentul acestui algoritm se poate observa în figura 8.

XGBoost se remarcă ca o implementare open-source și eficientă a algoritmului gradient boosted tree. Gradient boosting, o tehnică de învățare supervizată, își propune să prezică o variabilă țintă cu precizie prin amalgamarea estimărilor dintr-o colecție de modele mai simple și mai slabe [1].

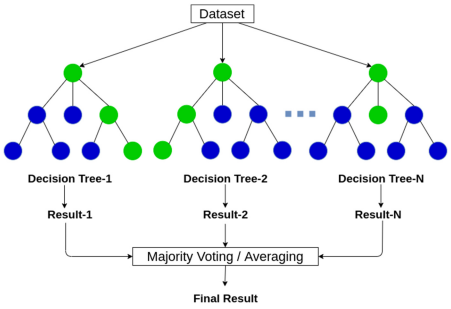

Figure 8: Random forest

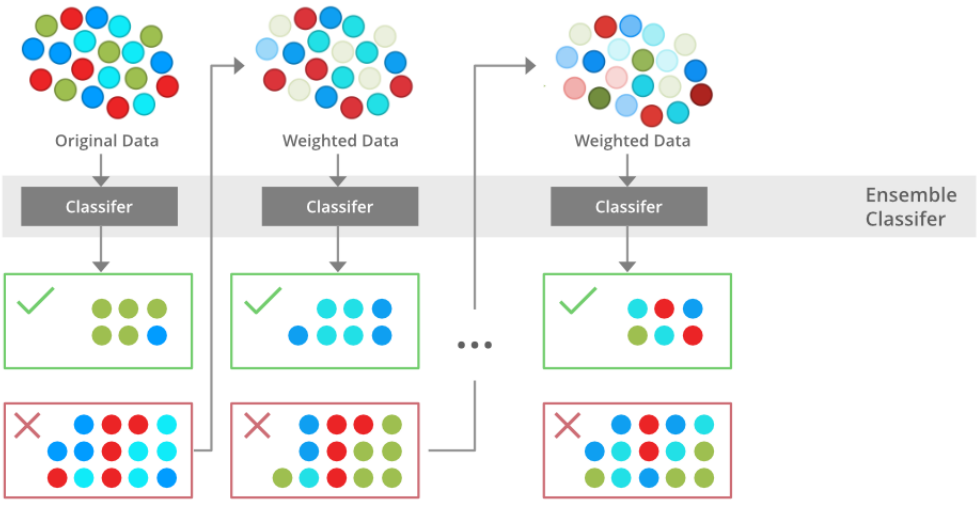

Figure 9: XGBoost

4.4 Analiză

Cu aceste unelte am propus un calcul al perioadei de timp în care o persoană a consumat ceva. Aceaste perioade de consum au fost urmatoarele: 'CL0': 'Niciodată folosit', 'CL1': 'Folosit acum mai mult de un deceniu','CL2': 'Folosit în ultimul deceniu', 'CL3': 'Folosit în ultimul an', 'CL4': 'Folosit în ultima lună', 'CL5': 'Folosit în ultima săptămână', 'CL6': 'Folosit în ultima zi', fiecare având coduri reprezentative în baza de date.

Pentru analiza modelelor am evaluat performanța acestora prin acuratețe și precizie, fiecare model obținând scoruri destul de apropiate. Antrenarea modelelor s-a realizat pe același set de date, cu o proportie de 80% pentru datele de antrenare si 20% pentru datele de testare.

În urma acestei antrenări am observat că predicția pentru drogurile consumate iregulat de persoanele aflate in caz, obținea scoruri destul de mici comparativ cu cele unde consumul era regulat.

5 Testare și validare

5.1 Antrenarea modelului final

Ca model final am ales regresia logistică, acesta obținând cele mai bune scoruri. Modelul a fost inițializat cu "LogisticRegression" care face parte din librăria sklearn.linear_model. Atributele care s-au luat în considerare pentru clasificarea și predicția tipului de drog au fost Age, Gender, Education, Country, Ethnicity, Nscore, Escore, Oscore, AScore, Cscore, Impulsive, SS. Pentru predicția consumului de canabis am obținut o acuratețe de 42%, pentru nicotină 37%, iar pentru heroină 85%, ceea ce ne confirmă că un consum iregulat în rândul oamenilor scade acuratețea predicției.

5.2 Validare

În figura 10 se poate observa că scorurile de antrenare și de validare converg pe masură ce setul de date crește, ceea ce poate indica faptul că adăugarea mai multod date poate ajuta modelul să învețe mai bine. Pentru a valida în mod corespunzător modelul și a evita orice bias în seturile de date de antrenament și testare, am utilizat validarea încrucișată cu k-folds. Aceasta tehnică împarte setul de date în k sub-seturi și antrenează modelul de k ori, de fiecare dată utilizând un sub-set diferit pentru validare și restul pentru antrenament. Rezultatele obținute prin această metodă sunt mediate pentru a oferi o estimare mai robustă a performanței modelului.

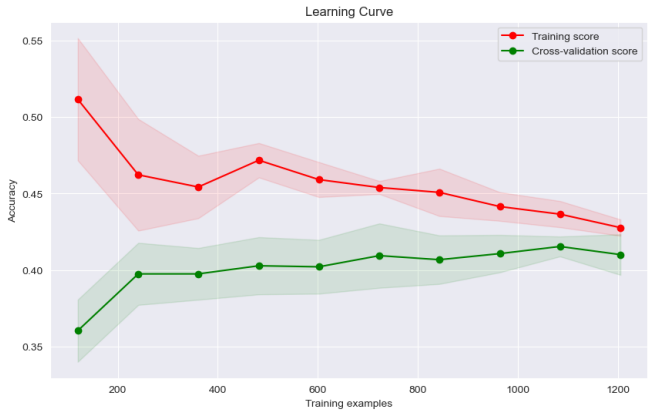

Figure 10: Curba de învățare

6 Rezultate

6.1 Testare

Pentru testarea modelului final am ales o abordare de testare manuală. Datele introduse au fost asemanătoare cu cele deja existente în baza de date, pentru a se putea face o predicție cât mai bună, totodată unele dintre acestea având mici modificări. Am testat aceste date pentru fiecare tip de drog în parte iar rezultatele obținute, comparativ cu datele inițiale, au fost foarte apropiate. Cu toate acestea impactul tipului de drog pe care s-a făcut clasificarea și predicția a fost foarte mare, lucru care se observă în obținerea rezultatelor.

6.2 Interpretarea rezultatelor

Chiar dacă rezultatele obținute nu au fost exacte, ținând cont că s-a făcut o predicție a unei perioade în care o persoană ar cunsuma un anumit tip de drog si nu calcularea unui număr exact, valorile acestora au fost foarte aproape de adevăr.

Din matricea de confuzie prezentată în figura 11 se poate observa următoarele caracteristici: pe diagonala principală clasele 0 și 1 au un număr relativ mare de instanțe corect clasificate (61 și 81 respectiv), ceea ce indică o performanță bună pentru aceste clase; clasa 2 are multe instanțe clasificate incorect ca fiind din clasa 0 și 1 (30 și 14 respectiv). clasele 3, 4, 5 și 6 au majoritatea instanțelor clasificate incorect, ceea ce indică o dificultate a modelului în a diferenția aceste clase; clasele 5 și 6 sunt frecvent confundate cu clasa 1, ceea ce sugerează că modelul are dificultăți în a distinge între aceste clase.

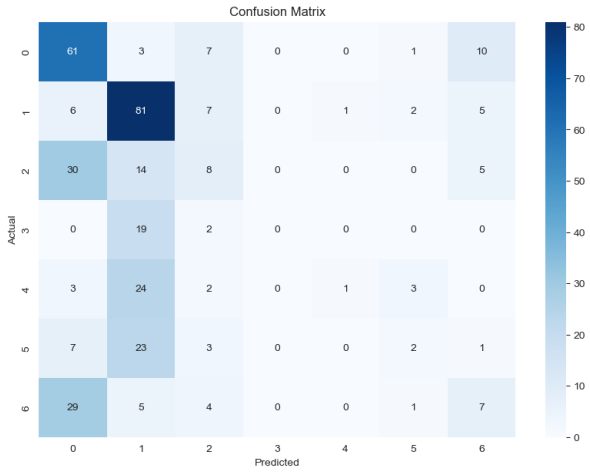

Figure 11: Matricea de confuzie

7 Concluzii

În cadrul acestui proiect, am investigat și analizat datele referitoare la consumul de droguri din baza de date "DrugConsumptionQuantified", cu scopul de a înțelege mai bine factorii care influențează comportamentul de consum de droguri și de a dezvolta strategii eficiente de prevenire și intervenție.

Prin intermediul analizei noastre, am identificat o serie de factori demografici, de personalitate și de mediu care sunt asociate cu consumul de droguri. Am constatat că vârsta, genul, educația și anumite trăsături de personalitate au o influență semnificativă asupra comportamentului de consum de droguri, în timp ce factorii sociali și culturali pot juca, de asemenea, un rol important. Utilizând diverse tehnici de analiză a datelor și de modelare predictivă, am dezvoltat modele precise și robuste pentru anticiparea comportamentului de consum de droguri în diferite contexte și populații. Aceste modele pot fi utilizate pentru a identifica persoanele cu risc crescut de consum de droguri și pentru a ghida intervențiile și politica de prevenire în mod eficient.

În concluzie, rezultatele noastre aduc o contribuție semnificativă la înțelegerea și abordarea consumului de droguri, evidențiind importanța factorilor individuali și sociali în determinarea acestui comportament complex. Prin integrarea aspectelor teoretice și a analizei practice a datelor, ne-am propus să oferim insight-uri utile pentru dezvoltarea de politici și intervenții eficiente în combaterea consumului de droguri și promovarea sănătății și bunăstării societății în ansamblu. Este crucial ca eforturile noastre să continue în acest domeniu, iar rezultatele noastre să fie aplicate în practică pentru a aduce schimbări pozitive și durabile în comunitățile noastre.

References

[1] Zeliha Ergul Aydin and Zehra Kamisli Ozturk. "Performance analysis of XGBoost classifier with missing data". In: *Manchester Journal of Artificial Intelligence and Applied Sciences (MJAIAS)* 2.02 (2021), p. 2021.

[2] Nils Braakmann and Simon Jones. "Cannabis depenalisation, drug consumption and crime–Evidence from the 2004 cannabis declassification in the UK". In: *Social Science & Medicine* 115 (2014), pp. 29–37.

[3] Joseph L Gastwirth. "The estimation of the Lorenz curve and Gini index". In: *The review of economics and statistics* (1972), pp. 306–316.

[4] Mark A Hall. "Correlation-based feature selection for machine learning". PhD thesis. The University of Waikato, 1999.

[5] Michael P LaValley. "Logistic regression". In: *Circulation* 117.18 (2008), pp. 2395–2399.

[6] Cristina Lopez-Mayan and Catia Nicodemo. ""If my buddies use drugs, will I?" Peer effects on Substance Consumption Among Teenagers". In: *Economics & Human Biology* 50 (2023), p. 101246.

[7] Christian Lüscher and Mark A Ungless. "The mechanistic classification of addictive drugs". In: *PLoS medicine* 3.11 (2006), e437.

[8] Howard Parker. "How young Britons obtain their drugs: Drugs transactions at the point of consumption". In: *Crime Prevention Studies* 11 (2000), pp. 59–82.

[9] Steven J Rigatti. "Random forest". In: *Journal of Insurance Medicine* 47.1 (2017), pp. 31–39.

[10] Claude Elwood Shannon. "A mathematical theory of communication". In: *ACM SIGMOBILE mobile computing and communications review* 5.1 (2001), pp. 3–55.