**1. Definirea problemei**

Datasetul cuprinde informații despre studenți colectate la momentul înscrierii. Acestea includ detalii despre parcursul academic, demografie și factori socio-economici. Informațiile sunt variate, acoperind starea civilă, modul de aplicare la universitate, cursurile urmate, prezența la cursuri (diurn/seral), calificările anterioare, naționalitatea, ocupația părinților și multe altele. Toate aceste date sunt esențiale pentru a înțelege contextul fiecărui student și a modela potențialele riscuri asociate parcursului său academic.

Problema centrală a acestui proiect este identificarea timpurie a studenților cu risc de abandon universitar, cu scopul de a reduce ratele de abandon și eșec academic în învățământul superior. Se utilizează un set de date complex, care include informații despre parcursul academic, demografie și factori socio-economici ai studenților la momentul înscrierii. Obiectivul este de a clasifica studenții în trei categorii la sfârșitul duratei normale de studiu: abandon, înmatriculare și absolvire.

**2. Informații necesare pentru rezolvare**

Pentru a aborda această problemă, sunt necesare următoarele informații din setul de date:

* Statutul marital, modul de aplicare la universitate, cursurile urmate, prezența la cursuri, calificările anterioare, naționalitatea și ocupația părinților.
* Date despre performanța academică, cum ar fi notele obținute și numărul de unități curriculare evaluate, înscrieri și aprobări.
* Date demografice și socio-economice, cum ar fi vârsta la înscriere, genul, dacă studentul este deplasat și dacă are nevoi educaționale speciale.

**3. Metode de calcul, algoritmi, formule de calcul utilizate**

***Calculul Ratei de Succes:***

Calculul ratei de succes pentru primul semestru se face folosind formula:

Acest calcul este aplicat fiecărui student folosind funcția *apply().*

**Algoritmi Utilizați:**

Pentru predicția categoriei de rezultat a studentului (abandon, înmatriculare, absolvire) se folosește algoritmul RandomForestClassifier din biblioteca scikit-learn, un algoritm bazat pe învățarea ansamblului de arbori de decizie.

**4. Prezentarea rezultatelor**

**Cerinte python**

* utilizarea listelor și a dicționarelor, incluzând metode specifice acestora;
* definirea și apelarea unor funcții
* utilizarea structurilor condiționale;
* importul unui fișier csv sau json în pachetul pandas;
* accesarea datelor cu loc și iloc;
* utilizarea funcțiilor de grup;
* prelucrări statistice, gruparea și agregarea datelor în pachetul pandas;
* reprezentarea grafică a datelor cu pachetul matplotlib;
* utilizarea pachetului scikit-learn;

**Prelucrarea datelor**

***Importarea bibliotecilor necesare***

A computer code with white text

Description automatically generated

***Citirea datelor și vizualizarea datelor***

A screenshot of a computer screen

Description automatically generated

**Analiza datelor**

1. Calcularea ratei de succes pe primul semestru

Definirea problemei: Problema definită este calcularea ratei de succes a studenților în primul semestru, care este esențială pentru a evalua performanța academică inițială și pentru a identifica studenții care pot avea nevoie de sprijin suplimentar pentru a preveni abandonul școlar. Rata de succes este definită ca raportul dintre numărul de unități curriculare aprobate și numărul de unități curriculare la care studentul s-a înscris în primul semestru. Această măsură permite instituțiilor să monitorizeze eficacitatea programelor lor educaționale și să intervina în timp util pentru a ajuta studenții care nu își ating potențialul complet.

A screenshot of a computer program

Description automatically generated

2. Maparea statutului marital, genului și grupei de vârstă

Descrierea problemei: problema specificată aici este îmbunătățirea ușurinței de interpretare a datelor setului de date al studenților prin transformarea valorilor numerice ale variabilelor categorice în etichete textuale mai ușor de înțeles. Această transformare ajută la analiza vizuală și statistică a datelor, permițând analiștilor să înțeleagă mai bine distribuția caracteristicilor cum ar fi starea civilă, genul și grupa de vârstă a studenților.

A screenshot of a computer program

Description automatically generated

3. Vizualizarea distribuției rezultatelor studenților în funcție de grupa de vârstă și starea civila.

Descrierea problemei: problema adresată în acest segment al codului este vizualizarea distribuției rezultatelor studenților (cum ar fi abandon, înmatriculare, și absolvire) în funcție de două variabile demografice: grupa de vârstă și starea civilă. Scopul este de a identifica tendințe sau modele în date care ar putea sugera corelații între aceste caracteristici demografice și succesul academic al studenților.

A screenshot of a computer screen

Description automatically generated

4. Dezvoltarea unui model de învățare automată pentru a clasifica studenții în categorii.

Descrierea problemei: problema abordată în acest cod implică dezvoltarea unui model de învățare automată pentru a clasifica studenții în trei categorii posibile bazate pe rezultatele lor academice: abandon (Dropout), înmatriculare continuă (Enrolled) și absolvire (Graduate). Scopul este de a utiliza datele despre studenți, cum ar fi statutul marital, genul, modul de aplicare, cursul, naționalitatea, calificările și ocupația părinților, grupa de vârstă și rata de succes din primul semestru pentru a prezice aceste rezultate. Acest model ajută instituțiile de învățământ să identifice studenții cu risc ridicat de abandon și să intervină în mod corespunzător.

A screen shot of a computer program

Description automatically generated

**Interpretare economică**

Rezultatele raportului de clasificare arată următoarele:

* Abandon (Dropout): Precizia de 83.4% și rechemarea de 78.1% indică o capacitate bună a modelului de a identifica corect studenții care vor abandona, cu un risc relativ scăzut de a clasifica greșit studenți care nu vor abandona.
* Înmatriculare (Enrolled): Precizia scăzută de 56.7% și rechemarea de 30.4% sugerează că modelul are dificultăți în a identifica corect studenții care continuă studiile fără a absolvii sau abandona în perioada evaluată.
* Absolvire (Graduate): Precizia de 77.4% și rechemarea de 93.7% indică o capacitate excelentă a modelului de a identifica studenții care absolvă, cu un număr foarte mic de absolviri ratate.

**Cerințe SAS**

* Crearea unui set de date SAS din fișiere externe;
* Crearea și folosirea de formate definite de utilizator;
* Procesarea iterativă și condițională a datelor;
* Crearea de subseturi de date;
* Utilizarea de funcții SAS;
* Combinarea seturilor de date prin proceduri specifice SQL;
* Utilizarea de masive;
* Utilizarea de proceduri pentru rapoarte;
* Folosirea de proceduri statistice;
* Generarea de grafice;

**Definirea problemei**

Scopul acestei secțiuni SAS a proiectului este de a procesa și analiza un set de date privind performanța studenților, pentru a identifica factorii care influențează succesul academic. Analiza se concentrează pe corelațiile dintre variabilele demografice, academice și socio-economice și performanța studenților.

**Informații necesare pentru rezolvare**

Informațiile cheie folosite în analiza SAS includ:

* Statutul marital, rata șomajului, rata inflației și PIB-ul, care sunt combinate pentru a analiza influența factorilor externi asupra performanței studenților.
* Datele despre înscrierile și evaluările curriculare ale studenților pentru a evalua performanța academică.
* Detalii demografice cum ar fi vârsta și genul, care sunt folosite pentru a studia distribuția notelor de admitere.

**Metode de calcul, Algoritmi și Formule de Calcul Utilizate**

Procesarea datelor:

* Imputarea valorilor lipsă cu 0 pentru unitățile curriculare ale studenților, folosind un array și o buclă do.
* Combinarea seturilor de date pentru a include variabilele economice relevante în analiza performanței studenților.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

SQL și Join-uri:

* Utilizarea PROC SQL pentru a combina datele studentului cu factorii externi, filtrând pentru studenții peste 18 ani și de gen masculin, ordonând după nota de admitere.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Crearea de rapoarte:

* Utilizarea PROC REPORT pentru a genera rapoarte despre statutul marital, genul și rezultatele bursierilor în funcție de media notelor de admitere.

A screenshot of a computer code

Description automatically generated

A screenshot of a document

Description automatically generated

Generarea de grafice:

* PROC SGPLOT pentru a crea un scatter plot care arată distribuția notelor de admitere în funcție de vârstă la înscriere și gen.

A close up of words

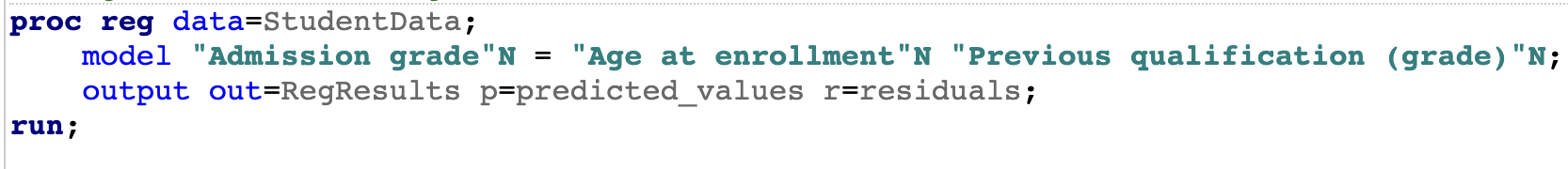
Description automatically generated

A screen shot of a graph

Description automatically generated

Analiza regresiei:

* PROC REG pentru a modela relația dintre vârsta la înscriere și calificarea anterioară asupra notei de admitere, producând valorile prezise și reziduurile.



A screenshot of a computer screen

Description automatically generated

**Interpretare economică**

Analiza economică a rezultatelor subliniază cum factorii socio-economici externi, cum ar fi rata șomajului și PIB-ul, influențează performanța academică. Prin identificarea și înțelegerea acestor corelații, instituțiile educaționale pot dezvolta strategii mai bine targetate pentru a sprijini studenții vulnerabili, optimizând utilizarea resurselor pentru a îmbunătăți rezultatele educaționale și pentru a minimiza riscul de abandon școlar.