Student-Performance

Darius Avram

May 2025

Contents

1 Introducere

1.1 Scopul proiectului

Scopul acestui proiect a fost de a determina factorii decisivi care influențează performanța studenților. O astfel de analiză oferă nu doar o înțelegere mai profundă a performanței școlare, ci și un sprijin real în procesul educațional pentru profesori și părinți.

Prin utilizarea unor tehnici moderne de analiză a datelor, inclusiv modele de regresie și vizualizări statistice, scopul este de a construi o imagine clară asupra impactului anumitor factori asupra rezultatelor obținute de elevi. Rezultatele proiectului pot contribui la dezvoltarea unor strategii educaționale personalizate, eficiente și bazate pe date reale.

1.2 Motivația alegerii temei și datasetului

Am ales această temă deoarece consider că este important să știm ce factori ar putea influența negativ sau pozitiv procesul de învățare al studenților.

În prezent, sistemele educaționale se confruntă cu numeroase provocări: diferențe de oportunitate între elevi, lipsa resurselor adaptate nevoilor fiecăruia și dificultăți în adaptarea metodelor de predare la generațiile actuale. Aceste probleme nu pot fi abordate eficient fără o înțelegere clară și obiectivă a factorilor care modelează parcursul școlar.

Cu ajutorul rezultatelor care reies în urma acestui proiect, putem schimba felul în care abordăm predarea materiilor (stilul de predare, timpul alocat fiecărei sesiuni de învățare, ore remote vs fizic etc).

De asemenea, utilizarea unui dataset public și bine documentat asigură replicabilitatea rezultatelor și permite compararea cu alte studii similare.

2 Contextul Proiectului și Cerințe

2.1 Descrierea datasetului

Datasetul[4] pe care l-am folosit pentru acest proiect provine de pe platforma Kaggle.
Acesta include informații despre **6378** de elevi pe **18** coloane relevante:

- număr de ore de studiu pe zi
- prezență
- nivelul de implicare al părinților
- acces la resurse
- activități extracuriculare
- ore de somn
- note anterioare(o medie)
- motivație
- acces la internet
- ore suplimentare
- venit total al familiei
- calitatea prfesorilor
- tipul școlii(public/privat)
- influența colegilor
- activități fizice
- disabilități de învățare
- nivelul de educație al părinților
- distanța față de casă
- gen

2.2 Obiectivele analizei

Scopul principal al acestei analize este să înțelegem mai bine cum diferite caracteristici ale elevilor (genul, mediul din care provin sau dacă au parte de ajutor suplimentar la învățătură etc) le influențează rezultatele la școală. Analizând datele disponibile, proiectul își propune să descopere legături importante între acești factori și notele obținute.

Pe lângă această parte descriptivă, proiectul încearcă și să construiască modele care pot prezice ce note ar putea obține elevii în funcție de anumite informații despre ei. Asta ne ajută nu doar să înțelegem mai bine ce contează în procesul de învățare, ci și să oferim soluții concrete pentru îmbunătățirea performanței. Pentru a face rezultatele cât mai clare și ușor de înțeles, analiza este însoțită de grafice și teste care verifică dacă concluziile trase sunt corecte și de încredere.

Pentru acest priect am propus o serie de obiective clare:

- Identificarea principalilor factori care influențează performanța școlară a elevilor
- Analiza corelațiilor dintre variabilele independente (gen, nivelul de educație al părinților, participarea la cursuri etc.)

- Detectarea dezechilibrelor între diferite categorii de elevi (de exemplu în funcție de gen)
- Construirea unor modele predictive capabile să estimeze scorurile elevilor pe baza caracteristicilor lor
- Vizualizarea datelor pentru a evidenția tipare și relații relevante, folosind grafice intuitive și ușor de interpretat
- Validarea modelelor utilizate, pentru a asigura acuratețea concluziilor

3 Aspecte Teoretice și State-of-the-Art

3.1 Factori relevanți pentru performanța școlară

Performanța școlară este influențată de o varietate de factori care pot fi grupați în mai multe categorii: factori individuali, de mediu și instituționali. Factorii individuali includ aptitudinile cognitive, motivația, stilul de învățare și starea psihologică a elevului[2]. De exemplu, motivația intrinsecă și nivelul de autoeficacitate au fost corelate cu performanțe academice superioare[5].

Factorii de mediu se referă la influența familiei, a mediului social și economic, precum și a grupurilor de colegi[10]. Sprijinul parental[9] și nivelul de educație al părinților sunt indicatori puternici ai succesului școlar[8].

În ceea ce privește factorii instituționali, calitatea procesului didactic, dotările școlii și metodele pedagogice utilizate joacă un rol esențial[1]. De asemenea, frecvența și implicarea în activități extracurriculare contribuie la dezvoltarea competențelor[6] transversale și la creșterea performanței academice.

Cercetările recente indică faptul că performanța elevilor este determinată printr-un complex de interacțiuni între acești factori.

3.2 Metode de analiză și clasificare

Analiza performanței școlare utilizează o gamă largă de metode statistice și algoritmi de învățare automată pentru a identifica tipare și a prezice rezultatele academice.

Metodele statistice clasice includ analiza descriptivă, analiza de corelație și regresia liniară sau logistică. Acestea oferă o bază pentru înțelegerea relațiilor între variabile, însă sunt limitate[3] în captarea relațiilor complexe și non-liniare.

În ultimii ani, metodele de învățare automată au fost intens utilizate datorită capacității lor de a procesa volume mari de date și de a învăța din exemple. Algoritmi precum arborii decizionali, rețelele neuronale artificiale și metodele ensemble (de exemplu Random Forest) oferă rezultate superioare în sarcini de clasificare și predicție.[7]

Pe lângă performanța predictivă, interpretabilitatea modelelor este o preocupare importantă, mai ales în domeniul educațional, unde deciziile trebuie să fie explicabile și transparente. Astfel, metode precum regresia logistică și arborii decizionali sunt adesea preferate în contexte practice.

Mai recent, tehnici avansate de procesare a limbajului natural (NLP) și analiza sentimentelor sunt folosite pentru a evalua feedback-ul textual al elevilor și pentru a înțelege mai bine factorii emoționali ce influențează învățarea.[11]

În concluzie, combinarea metodelor statistice tradiționale cu algoritmi moderni de învățare automată oferă un cadru solid pentru analiza și îmbunătățirea performanței școlare.

4 Implementarea

4.1 Preprocesarea datelor

Preprocesarea datelor reprezintă o etapă esențială în pregătirea datelor pentru modelele de învățare automată, asigurând transformarea valorilor calitative în formate numerice ce pot fi interpretate de algoritmi.

În etapa inițială a proiectului, am convertit manual în valori numerice întregi toate coloanele care aveau stringuri, prin crearea unor dicționare de mapare specifice fiecărei variabile. Această abordare m-a ajutat să înțeleg mai ușor datele codificate:

```
- Parental_Involvement (low - 0, medium - 1, high - 2)
- Access_to_Resources (low - 0, medium - 1, high - 2)
- Extracurricular_Activities (no - 0, yes - 1)
- Motivation_Level (low - 0, medium - 1, high - 2)
- Internet_Access (no - 0, yes - 1)
- Family_Income (low - 0, medium - 1, high - 2)
- Teacher_Quality (low - 0, medium - 1, high - 2)
- School_Type (Public - 0, Private - 1)
- Peer_Influence (negative - 0, neutral - 1, positive- 2)
- Learning_Disabilities (no - 0, yes - 1)
- Parental_Education_Level (high school - 0, college - 1, postgraduate - 2)
- Distance_from_Home (near - 0, moderate - 1, far - 2)
- Gender (male - 0, female - 1)
```

După ce am transformat inițial stringurile în numere, pentru a evita ca modelele să interpreteze aceste numere ca având o ordine între ele, am folosit *one-hot encoding*. Aceasta înseamnă că fiecare categorie a devenit o coloană separată, cu valoare 0 sau 1, astfel încât modelul să nu creadă că o categorie ar avea o importanță mai mare sau mai mică decât alta.

Folosind această metodă, datele au devenit potrivite pentru majoritatea algoritmilor de clasificare și regresie. Pentru aplicarea *one-hot encoding*, am folosit funcția get_dummies() disponibilă în biblioteca `pandas`.

Astfel, preprocesarea datelor a fost făcută în două etape: mai întâi am mapat manual categoriile în numere pentru o mai bună înțelegere, apoi am aplicat *one-hot encoding* pentru a obține o reprezentare mai clară și mai ușor de folosit de către algoritmi.

4.2 Modele folosite

Pentru a înțelege și a estima modul în care anumiți factori influențează performanța școlară, am folosit mai multe modele. Cele trei modele aplicate în cadrul proiectului sunt: Logistic Regression, Decision Tree și Random Forest.

Logistic Regression

Logistic Regression este un model statistic folosit pentru a prezice probabilitatea ca o observație să aparțină uneia dintre două sau mai multe clase. În contextul acestui proiect, a fost utilizată pentru a evalua probabilitatea ca un elev să obțină un scor ridicat sau scăzut în funcție de caracteristicile sale (precum genul, nivelul de educație al părinților, etc.). Modelul are avantajul de a fi interpretabil și eficient pentru date liniare, fiind util pentru a evidenția ce factori au o influență mai mare asupra rezultatului.

Decision Tree

Decision Tree este un model intuitiv, bazat pe reguli, care împarte setul de date în funcție de anumite condiții până ajunge la o predicție finală. Fiecare nod intern al arborelui corespunde unui test pe o caracteristică, iar frunzele reprezintă rezultatele. În acest proiect, Decision Tree a fost folosit pentru a înțelege mai clar cum combinațiile de factori pot duce la diferite performanțe academice. Deși poate fi predispus la overfitting, modelul oferă o vizualizare clară a deciziilor și este ușor de interpretat.

Random Forest

Random Forest este un model care combină mai mulți arbori de decizie pentru a îmbunătăți acuratețea predicțiilor. Fiecare arbore este antrenat pe un subset aleatoriu al datelor, iar rezultatul final este obținut prin votul majoritar (în clasificare) sau media predicțiilor (în regresie). Random Forest are avantajul de a reduce riscul de overfitting și de a oferi o performanță mai stabilă decât un decision tree simplu. În cadrul acestui proiect, modelul a fost utilizat pentru a construi predicții robuste și pentru a analiza importanța variabilelor.

5 Testare și validare

În acest capitol voi descrie metodele prin care au fost evaluate modelele de învățare automată utilizate în cadrul proiectului. Scopul este de a verifica cât de bine generalizează modelele antrenate pe date noi și necunoscute.

5.1 Împărțirea datelor în training/test

Pentru a evalua performanța modelelor într-un mod obiectiv, am împărțit datele în două subseturi: 80% pentru antrenare (training) și 20% pentru testare (test). Împărțirea s-a realizat cu ajutorul funcției train_test_split din biblioteca sklearn.model_selection, folosind o valoare fixă pentru random_state pentru a asigura reproductibilitatea rezultatelor.

În preprocesare, datele numerice au fost normalizate folosind StandardScaler, iar datele categorice au fost codificate cu LabelEncoder. Aceste etape au contribuit la creșterea performanței și stabilității modelelor.

5.2 Confusion Matrix

Confusion Matrix este un tabel utilizat pentru a evalua performanța unui algoritm de clasificare. Aceasta oferă o reprezentare clară a modului în care modelul face predicțiile, comparând valorile prezise de model cu cele reale din setul de date de testare.

- Matricea de confuzie conține patru elemente principale în cazul unei clasificări binare:
- True Positives (TP) – cazuri în care modelul a prezis corect clasa pozitivă;
- True Negatives (TN) – cazuri în care modelul a prezis corect clasa negativă;
- False Positives (FP) – cazuri în care modelul a prezis greșit clasa pozitivă (defapt era negativă);
- False Negatives (FN) – cazuri în care modelul a prezis greșit clasa negativă (defapt era pozitivă).

Acest tabel permite evaluarea unor metrici importante precum:

- Acuratețea: propor'ie predicții corecte,
- Precizia: cât de precise sunt predicțiile pozitive,
- Recall-ul: cât de bine identifică modelul toate cazurile pozitive,
- F1-score: media dintre precizie și recall, utilă în cazul claselor dezechilibrate.

În cadrul proiectului, pentru fiecare model antrenat (Logistic Regression, Decision Tree, Random Forest), am generat o matrice de confuzie pe datele de test, afișată sub formă de heatmap. Acest lucru a permis o analiză vizuală intuitivă a erorilor făcute de model, facilitând comparația între diferitele modele și identificarea celor care au avut cea mai bună capacitate de clasificare.

6 Rezultate

Pentru a evalua performanța modelelor de clasificare utilizate în acest proiect, am folosit două metode esențiale: Cross-Validation(scorul mediu obținut prin validare încrucișată) și Model Comparison (acuratețea pe setul de test). Rezultatele obținute sunt sintetizate în tabelul de mai jos:

```
--- Performing Cross-Validation ---
Logistic Regression Cross-Validation Score: 0.8336 ± 0.0059
Decision Tree Cross-Validation Score: 0.2103 ± 0.0055
Random Forest Cross-Validation Score: 0.2577 ± 0.0068

--- Model Comparison ---
                 Model  Test Accuracy  CV Accuracy
0  Logistic Regression       0.827586     0.833593
2        Random Forest       0.268809     0.257741
1        Decision Tree       0.205329     0.210309
```

Modelul de Logistic Regression a avut cele mai bune rezultate, obținând o acuratețe de 82.76% pe setul de test și o acuratețe medie de 83.36% în urma Cross-Valida, cu o deviație standard redusă. Acest lucru sugerează că modelul este stabil și generalizează bine pe date care nu apar]n setul de antrenament.

În schimb, modelele bazate pe arbori de decizie (Decision Tree și Random Forest) au înregistrat performanțe semnificativ mai slabe. Acuratețea testului a fost sub 30% în ambele cazuri, iar scorurile de cross-validation au fost de asemenea scăzute. Acest comportament poate fi explicat prin:

- posibilă supraînvățare (overfitting) pe setul de antrenament
- lipsa unui număr suficient de estimatori sau adâncime corespunzătoare în cazul Random Forest
- distribuția inegală a claselor în datele de ieșire (Exam_Score), ceea ce ar putea afecta sensibil modelele bazate pe arbori.

7 Concluzii

Pe baza acestor rezultate, regresia logistică este modelul cel mai potrivit pentru problema de clasificare analizată în acest proiect. Acesta va fi utilizat în etapa finală pentru realizarea predicțiilor asupra datelor noi și extragerea concluziilor relevante.

7.1 Factori influenți

Pentru a identifica ce variabile influențează cel mai mult rezultatul obținut la examen (Exam_Score), am antrenat un model **Random Forest Regressor** pe întregul set de date. Acest model permite extragerea importanței fiecărei variabile de intrare în funcție de contribuția sa în predicție.

Caracteristică	**Importanță**
Attendance	0.3847
Hours_Studied	0.2425
Previous_Scores	0.0818
Parental_Involvement	0.0376
Tutoring_Sessions	0.0353
Access_to_Resources	0.0328
Sleep_Hours	0.0269
Physical_Activity	0.0265
Family_Income	0.0185
Parental_Education_Level	0.0183
Peer_Influence	0.0161
Distance_from_Home	0.0151
Motivation_Level	0.0145
Teacher_Quality	0.0136
Learning_Disabilities	0.0091
School_Type	0.0077
Extracurricular_Activities	0.0072
Internet_Access	0.0063
Gender	0.0057

Se observă că cei mai influenți factori sunt:

- **Attendance (prezența la cursuri)** – cu o importanță de peste 38%, fiind cel mai puternic predictor al performanței academice;
- **Hours_Studied (orele de studiu)** – contribuie în proporție de 24%;
- **Previous_Scores (notele anterioare)** – reflectă performanța anterioară a studentului și au o influență semnificativă;

References

[1] Sarah J Carrington Alejandra Espinosa Andrade, León Padilla. Educational spaces: The relation between school infrastructure and learning outcomes. 2024. (Link).

[2] Mitra Amini Nasrin Shokrpour Ali Asghar Hayat, Karim Shateri. Relationships between academic self-efficacy, learning-related emotions, and metacognitive learning strategies with academic performance in medical students: a structural equation model. 2020. (Link).

[3] Trevor Hastie Robert Tibshirani Gareth James, Daniela Witten. An introduction to statistical learning. 2013. (Link).

[4] Lai Ng. Student performance dataset. 2024. (Dataset Link).

[5] Shaiful Annuar Khalid Norshimah Rahman. Academic self-efficacy, intrinsic motivation and academic achievement: Moderating effect of gender. 2023. (Link).

[6] Jhalak Sharma. Importance of extracurricular activities in student development. 2025. (Link).

[7] Jan Stihec. Random forests in machine learning for advanced decision-making. 2014. (Link).

[8] Ram Bahadur Bhandari Tatwa Prasad Timsina. Examining the impact of parents' education on students' academic achievements. 2024. (Link).

[9] Positive Action Team. The impact of parental involvement: Statistics on academic success. 2023. (Link).

[10] Quan Lu Chao Zhang Ruoxi Li Yanhong Shao, Shumin Kang. How peer relationships affect academic achievement among junior high school students: The chain mediating roles of learning motivation and learning engagement. 2024. (Link).

[11] Ali Shariq Imran Krenare Pireva Nuci Zenun Kastrati, Fisnik Dalipi and Ahmad Wani. Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. 2021. (Link).