Dataset: Prédire le décrochage et la réussite scolaire des élèves

Lien officiel : https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

1. Contexte et objectifs

Ce jeu de données accompagne l'étude « Early Prediction of Student's Performance in Higher Education: A Case Study » (Martins et al., 2021), qui vise à :

- Identifier **précocement** les étudiants à risque de décrochage ou d'échec académique.
- Proposer des mesures de soutien adaptées avant le début de l'année universitaire.
- Répartition des statuts des étudiants dans le dataset :
  - **Graduate** (Réussite) : 2209 (49.9 %)
  - **Dropout** (Echec) : 1421 (32.1 %)
  - **Enrolled** (Inscrit) : 794 (17.9 %)

Les données couvrent la période **2008/09–2018/19** et portent sur des étudiants de l'Institut Polytechnique de Portalegre (Portugal).

2. Caractéristiques des données initiales

- **Type** : Données tabulaires
- **Instances** : 4424 échantillons initiaux
- **Variables** : 36 initiales d'entrée (features) + 1 variable cible (target)

Variables principales

1. Marital Status
Statut marital de l'étudiant au moment de l'inscription.
Valeurs possibles :

- 1 : Célibataire
- 2 : Marié
- 3 : Veuf
- 4 : Divorcé
- 5 : Union libre (concubinage)
- 6 : Séparation légale

2. Application mode
Mode d'admission utilisé par l'étudiant.
Valeurs possibles :

- 1 : 1re phase, contingent général
- 2 : Arrêté n°612/93

- 5 : 1re phase, contingent spécial (îles Açores)
- 7 : Titulaires d'autres diplômes supérieurs
- 10 : Arrêté n°854-B/99
- 15 : Étudiant international (bachelor)
- 16 : 1re phase, contingent spécial (îles Madère)
- 17 : 2e phase, contingent général
- 18 : 3e phase, contingent général
- 26 : Arrêté n°533-A/99, item b2 (plan différent)
- 27 : Arrêté n°533-A/99, item b3 (autre institution)
- 39 : Âge supérieur à 23 ans
- 42 : Transfert
- 43 : Changement de cursus
- 44 : Titulaires de diplômes de spécialisation technologique
- 51 : Changement d'institution ou de cursus
- 53 : Titulaires de diplômes de court cycle
- 57 : Changement d'institution ou de cursus (international)

3. Application order
Ordre de préférence dans lequel l'étudiant a soumis sa candidature.
Valeurs possibles :

- 0 : Premier choix
- 1 : Deuxième choix
- 2 : Troisième choix
- ...
- 9 : Dixième choix (dernier choix)

4. Course
Code numérique représentant le cursus universitaire choisi.
Valeurs possibles :

- 33 : Technologies de production de biocarburants
- 171 : Animation et design multimédia
- 8014 : Service social (cours du soir)
- 9003 : Agronomie
- 9070 : Design de communication
- 9085 : Soins infirmiers vétérinaires
- 9119 : Génie informatique
- 9130 : Équiculture
- 9147 : Management
- 9238 : Service social
- 9254 : Tourisme
- 9500 : Soins infirmiers
- 9556 : Hygiène bucco-dentaire
- 9670 : Publicité et gestion marketing
- 9773 : Journalisme et communication
- 9853 : Éducation de base
- 9991 : Management (cours du soir)

5. Daytime/evening attendance
Indique si l'étudiant suit les cours en journée ou en soirée.
Valeurs possibles :

- 1 : Cours en journée
- 0 : Cours en soirée

6. Previous qualification
Niveau académique atteint avant l'inscription à l'université.
Valeurs possibles :

- 1 : Enseignement secondaire
- 2 : Diplôme supérieur de premier cycle (Licence/Bachelor)
- 3 : Diplôme supérieur de deuxième cycle (Master)
- 4 : Diplôme supérieur de troisième cycle (Doctorat)
- 5 : Formation professionnelle continue
- 6 : Autres formations supérieures
- 9 : 12e année d'études non complétée
- 10 : 11e année d'études non complétée
- 12 : Autres formations de 11e année
- 14 : 10e année
- 15 : 10e année non complétée
- 19 : Éducation de base 3e cycle (9e/10e/11e année)
- 38 : Éducation de base 2e cycle (6e/7e/8e année)
- 39 : Diplôme de spécialisation technologique
- 40 : Diplôme de premier cycle universitaire
- 42 : Diplôme technique professionnel supérieur
- 43 : Master (2e cycle)

7. Previous qualification (grade)
Note obtenue au diplôme précédent, sur une échelle allant de 0 à 200.
Valeurs possibles :

- Continu de 0.0 à 200.0

8. Nacionality
Nationalité de l'étudiant codée numériquement.
Valeurs possibles :

- 1 : Portugais
- 2 : Allemand
- 6 : Espagnol
- 11 : Italien
- 13 : Néerlandais
- 14 : Anglais
- 17 : Lituanien
- 21 : Angolais
- 22 : Cap-Verdien
- 24 : Guinéen

- 25 : Mozambicain
- 26 : Santoméen
- 32 : Turc
- 41 : Brésilien
- 62 : Roumain
- 100 : Moldave
- 101 : Mexicain
- 103 : Ukrainien
- 105 : Russe
- 108 : Cubain
- 109 : Colombien

9. Mother's qualification

Niveau d'éducation de la mère.

Valeurs possibles :

- 1 : Enseignement secondaire (12e année ou équivalent)
- 2 : Diplôme supérieur premier cycle
- 3 : Diplôme supérieur deuxième cycle
- 4 : Master (2e cycle)
- 5 : Doctorat (3e cycle)
- 6 : Formation continue
- 9 : 12e année non complétée
- 10 : 11e année non complétée
- 11 : 7e année (ancienne échelle)
- 12 : Autres formations 11e année
- 14 : 10e année
- 18 : Commerce général
- 19 : Éducation de base 3e cycle
- 22 : Formation technique professionnelle
- 26 : 7e année
- 27 : 2e cycle du lycée général
- 29 : 9e année non complétée
- 30 : 8e année
- 34 : Inconnu
- 35 : Analphabète (ne sait pas lire ou écrire)
- 36 : Sait lire sans formation de 4e année
- 37 : Éducation de base 1er cycle (4e/5e année)
- 38 : Éducation de base 2e cycle
- 39 : Cursus de spécialisation technologique
- 40 : Enseignement supérieur - diplôme (1er cycle)
- 41 : Cours d'études supérieures spécialisées
- 42 : Cours technique supérieur professionnel
- 43 : Enseignement supérieur - master (2e cycle)
- 44 : Enseignement supérieur - doctorat (3e cycle)

10. Father's qualification

Niveau d'éducation du père.

Valeurs possibles :

- 1 : Enseignement secondaire (12e année ou équivalent)
- 2 : Diplôme supérieur premier cycle
- 3 : Diplôme supérieur deuxième cycle
- 4 : Master (2e cycle)
- 5 : Doctorat (3e cycle)
- 6 : Formation continue
- 9 : 12e année non complétée
- 10 : 11e année non complétée
- 11 : 7e année (ancienne échelle)
- 12 : Autres formations 11e année
- 13 : 2e année complémentaire lycée
- 14 : 10e année
- 18 : Commerce général
- 19 : Éducation de base 3e cycle
- 20 : Formation complémentaire lycée
- 22 : Formation technique professionnelle
- 25 : Formation complémentaire non complétée
- 26 : 7e année
- 27 : 2e cycle du lycée général
- 29 : 9e année non complétée
- 30 : 8e année
- 31 : Cours généraux administration et commerce
- 33 : Comptabilité et administration
- 34 : Inconnu
- 35 : Analphabète
- 36 : Lecture sans formation complète
- 37 : Éducation de base 1er cycle
- 38 : Éducation de base 2e cycle
- 39 : Cursus de spécialisation technologique
- 40 : Enseignement supérieur - diplôme (1er cycle)
- 41 : Cours d'études supérieures spécialisées
- 42 : Cours technique supérieur professionnel
- 43 : Enseignement supérieur - master (2e cycle)
- 44 : Enseignement supérieur - doctorat (3e cycle)

11. Mother's occupation

Profession de la mère de l'étudiant, codée numériquement selon la classification nationale des professions.

Valeurs possibles :

- 0 : Étudiant
- 1 : Représentants du pouvoir législatif et des organes exécutifs, directeurs et cadres dirigeants
- 2 : Spécialistes des activités intellectuelles et scientifiques
- 3 : Techniciens et professions de niveau intermédiaire

- 4 : Personnel administratif
- 5 : Travailleurs des services personnels, de la sécurité et des ventes
- 6 : Agriculteurs et ouvriers qualifiés de l’agriculture, de la pêche et de la sylviculture
- 7 : Ouvriers qualifiés de l’industrie, de la construction et artisans
- 8 : Opérateurs d’installation et de machines et travailleurs de l’assemblage
- 9 : Travailleurs non qualifiés
- 10 : Professions des forces armées
- 90 : Autre situation
- 99 : (vide)
- 101 : Officiers des forces armées
- 102 : Sous-officiers des forces armées
- 103 : Autre personnel des forces armées
- 112 : Directeurs des services administratifs et commerciaux
- 114 : Directeurs de l’hôtellerie, de la restauration, du commerce et d’autres services
- 121 : Spécialistes des sciences physiques, des mathématiques, de l’ingénierie et des techniques connexes
- 122 : Professionnels de la santé
- 123 : Enseignants
- 124 : Spécialistes en finance, comptabilité, organisation administrative, relations publiques et commerciales
- 125 : Pas de correspondances
- 131 : Techniciens et professions intermédiaires des sciences et de l’ingénierie
- 132 : Techniciens et professionnels de santé de niveau intermédiaire
- 134 : Techniciens de niveau intermédiaire des services juridiques, sociaux, sportifs, culturels et similaires
- 135 : Techniciens en technologies de l’information et de la communication
- 141 : Employés de bureau, secrétaires et opérateurs de saisie de données
- 143 : Opérateurs de services de données, de comptabilité, statistiques, financiers et de registre
- 144 : Autre personnel de soutien administratif
- 151 : Travailleurs des services personnels
- 152 : Vendeurs
- 153 : Travailleurs des soins personnels et assimilés
- 154 : Personnel des services de protection et de sécurité
- 161 : Agriculteurs de culture de marché et ouvriers qualifiés de la production agricole et animale
- 163 : Agriculteurs, éleveurs, pêcheurs, chasseurs et cueilleurs de subsistance
- 171 : Ouvriers qualifiés du bâtiment et assimilés, à l’exception des électriciens
- 172 : Ouvriers qualifiés de la métallurgie, du travail des métaux et similaires
- 173 : Pas de correspondances
- 174 : Ouvriers qualifiés en électricité et en électronique
- 175 : Travailleurs de la transformation alimentaire, du bois, de l’habillement et autres industries et artisanats
- 181 : Opérateurs d’installations fixes et de machines
- 182 : Travailleurs de l’assemblage
- 183 : Conducteurs de véhicules et opérateurs d’équipements mobiles
- 191 : Pas de correspondances
- 192 : Travailleurs non qualifiés de l’agriculture, de la production animale, de la pêche et de la sylviculture

- 193 : Travailleurs non qualifiés de l'industrie extractive, de la construction, de la fabrication et du transport
- 194 : Aides à la préparation des repas
- 195 : Vendeurs ambulants (hors produits alimentaires) et prestataires de services de rue

12. Father's occupation
Profession du père de l'étudiant, codée numériquement selon la classification nationale des professions.
Valeurs possibles :

- 0 : Étudiant
- 1 : Représentants du pouvoir législatif et des organes exécutifs, directeurs et cadres dirigeants
- 2 : Spécialistes des activités intellectuelles et scientifiques
- 3 : Techniciens et professions de niveau intermédiaire
- 4 : Personnel administratif
- 5 : Travailleurs des services personnels, de la sécurité et des ventes
- 6 : Agriculteurs et ouvriers qualifiés de l'agriculture, de la pêche et de la sylviculture
- 7 : Ouvriers qualifiés de l'industrie, de la construction et artisans
- 8 : Opérateurs d'installation et de machines et travailleurs de l'assemblage
- 9 : Travailleurs non qualifiés
- 10 : Professions des forces armées
- 90 : Autre situation
- 99 : (vide)
- 101 : Officiers des forces armées
- 102 : Sous-officiers des forces armées
- 103 : Autre personnel des forces armées
- 112 : Directeurs des services administratifs et commerciaux
- 114 : Directeurs de l'hôtellerie, de la restauration, du commerce et d'autres services
- 121 : Spécialistes des sciences physiques, des mathématiques, de l'ingénierie et des techniques connexes
- 122 : Professionnels de la santé
- 123 : Enseignants
- 124 : Spécialistes en finance, comptabilité, organisation administrative, relations publiques et commerciales
- 125 : Pas de correspondances
- 131 : Techniciens et professions intermédiaires des sciences et de l'ingénierie
- 132 : Techniciens et professionnels de santé de niveau intermédiaire
- 134 : Techniciens de niveau intermédiaire des services juridiques, sociaux, sportifs, culturels et similaires
- 135 : Techniciens en technologies de l'information et de la communication
- 141 : Employés de bureau, secrétaires et opérateurs de saisie de données
- 143 : Opérateurs de services de données, de comptabilité, statistiques, financiers et de registre
- 144 : Autre personnel de soutien administratif
- 151 : Travailleurs des services personnels
- 152 : Vendeurs
- 153 : Travailleurs des soins personnels et assimilés
- 154 : Personnel des services de protection et de sécurité
- 161 : Agriculteurs de culture de marché et ouvriers qualifiés de la production agricole et animale
- 163 : Agriculteurs, éleveurs, pêcheurs, chasseurs et cueilleurs de subsistance

- 171 : Ouvriers qualifiés du bâtiment et assimilés, à l'exception des électriciens
- 172 : Ouvriers qualifiés de la métallurgie, du travail des métaux et similaires
- 173 : Pas de correspondances
- 174 : Ouvriers qualifiés en électricité et en électronique
- 175 : Travailleurs de la transformation alimentaire, du bois, de l'habillement et autres industries et artisanats
- 181 : Opérateurs d'installations fixes et de machines
- 182 : Travailleurs de l'assemblage
- 183 : Conducteurs de véhicules et opérateurs d'équipements mobiles
- 191 : Pas de correspondances
- 192 : Travailleurs non qualifiés de l'agriculture, de la production animale, de la pêche et de la sylviculture
- 193 : Travailleurs non qualifiés de l'industrie extractive, de la construction, de la fabrication et du transport
- 194 : Aides à la préparation des repas
- 195 : Vendeurs ambulants (hors produits alimentaires) et prestataires de services de rue

13. Admission grade
Note d'admission à l'université, calculée selon les critères nationaux portugais.
Valeurs possibles :

- Continu de 0.0 à 200.0

14. Displaced
Indique si l'étudiant a été déplacé de sa région d'origine pour ses études.
Valeurs possibles :

- 0 : Non (étudiant local)
- 1 : Oui (étudiant déplacé)

15. Educational special needs
Présence de besoins éducatifs spéciaux chez l'étudiant.
Valeurs possibles :

- 0 : Non
- 1 : Oui

16. Debtor
Statut de débiteur de l'étudiant envers l'institution.
Valeurs possibles :

- 0 : Non (pas de dettes)
- 1 : Oui (dettes en cours)

17. Tuition fees up to date
Indique si les frais de scolarité sont à jour.
Valeurs possibles :

- 0 : Non (frais en retard)
- 1 : Oui (frais à jour)

18. Gender
Sexe de l'étudiant.
Valeurs possibles :

- 1 : Homme
- 0 : Femme

19. Scholarship holder
Indique si l'étudiant bénéficie d'une bourse d'études.
Valeurs possibles :

- 0 : Non
- 1 : Oui

20. Age at enrollment
Âge de l'étudiant au moment de l'inscription.
Valeurs possibles :

- Entier ≥ 17

21. International
Statut d'étudiant international.
Valeurs possibles :

- 0 : Étudiant national
- 1 : Étudiant international

22. Curricular units 1st sem (credited)
Nombre de crédits ECTS acquis par équivalence au premier semestre (sans examen).
Valeurs possibles :

- Entier ≥ 0

23. Curricular units 1st sem (enrolled)
Nombre total d'unités d'enseignement auxquelles l'étudiant s'est inscrit au premier semestre.
Valeurs possibles :

- Entier ≥ 0

24. Curricular units 1st sem (evaluations)
Nombre d'examens passés au premier semestre.
Valeurs possibles :

- Entier ≥ 0

25. Curricular units 1st sem (approved)
Nombre de matières validées (réussies) au premier semestre.
Valeurs possibles :

- Entier ≥ 0

26. Curricular units 1st sem (grade)
Note moyenne obtenue dans les unités validées au premier semestre.
Valeurs possibles :

- Continu de 0.0 à 20.0 (échelle portugaise)

27. Curricular units 1st sem (without evaluations)
Nombre de matières inscrites au premier semestre pour lesquelles aucune évaluation n'a eu lieu.
Valeurs possibles :

- Entier ≥ 0

28. Curricular units 2nd sem (credited)
Nombre de crédits ECTS acquis par équivalence au deuxième semestre.
Valeurs possibles :

- Entier ≥ 0

29. Curricular units 2nd sem (enrolled)
Nombre total d'unités d'enseignement auxquelles l'étudiant s'est inscrit au deuxième semestre.
Valeurs possibles :

- Entier ≥ 0

30. Curricular units 2nd sem (evaluations)
Nombre d'examens passés au deuxième semestre.
Valeurs possibles :

- Entier ≥ 0

31. Curricular units 2nd sem (approved)
Nombre de matières validées (réussies) au deuxième semestre.
Valeurs possibles :

- Entier ≥ 0

32. Curricular units 2nd sem (grade)
Note moyenne obtenue dans les unités validées au deuxième semestre.
Valeurs possibles :

- Continu de 0.0 à 20.0 (échelle portugaise)

33. Curricular units 2nd sem (without evaluations)
Nombre de matières inscrites au deuxième semestre pour lesquelles aucune évaluation n'a eu lieu.
Valeurs possibles :

- Entier ≥ 0

34. Unemployment rate
Taux de chômage national (en pourcentage) lors de l'année d'inscription de l'étudiant.
Valeurs possibles :

- Continu

35. Inflation rate
Taux d'inflation national (en pourcentage) lors de l'année d'inscription.
Valeurs possibles :

- Continu

36. GDP
Produit intérieur brut par habitant du Portugal lors de l'année d'inscription, exprimé en unités relatives.
Valeurs possibles :

- Continu

Variable cible

37. Target
Classe à prédire représentant le statut académique final de l'étudiant.
Valeurs possibles :

- Graduate : Diplômé (réussite complète)
- Dropout : Décrochage scolaire
- Enrolled : Toujours inscrit (en cours d'études)

---

3. Features Rajouté :

Previous qualification Group – Classification Académique

Groupe 0 : Diplômes universitaires et post-universitaires

- 4 : Diplôme supérieur de troisième cycle (Doctorat)
- 43 : Master (2e cycle)
- 3 : Diplôme supérieur de deuxième cycle (Master)
- 2 : Diplôme supérieur de premier cycle (Licence/Bachelor)
- 40 : Diplôme de premier cycle universitaire
- 6 : Autres formations supérieures

Caractéristiques : Niveau d'études universitaire ou supérieure. Correspond au niveau le plus élevé d'accès direct à l'université. Les étudiants venant de ce groupe disposent déjà d'acquis solides, facilitant la réussite académique et l'accès aux cycles supérieurs.

Groupe 1 : Diplômes techniques, professionnels ou de spécialisation

- 42 : Diplôme technique professionnel supérieur
- 39 : Diplôme de spécialisation technologique
- 5 : Formation professionnelle continue

Caractéristiques : Viennent de filières techniques ou professionnelles reconnues et qualifiantes. Les étudiants de ce groupe ont de bonnes compétences pratiques et une orientation projet, mais leur réussite dans les

filières universitaires plus théoriques varie selon l'accompagnement.

Groupe 2 : Enseignement secondaire classique et parcours incomplets

- 1 : Enseignement secondaire
- 9 : 12e année d'études non complétée
- 10 : 11e année d'études non complétée
- 12 : Autres formations de 11e année
- 14 : 10e année
- 15 : 10e année non complétée

Caractéristiques : Fin de lycée (ou équivalent), parfois avec un parcours scolaire interrompu ou incomplet. Niveau de préparation académique variable ; nécessite souvent des dispositifs d'accompagnement ou de transition pour bien réussir à l'université.

Groupe 3 : Éducation de base ou parcours scolaires courts

- 19 : Éducation de base 3e cycle (9e/10e/11e année)
- 38 : Éducation de base 2e cycle (6e/7e/8e année)

Caractéristiques : Parcours scolaire interrompu précocement ou limité. Les étudiants issus de ce groupe sont considérés comme à risque accru d'échec ou d'abandon sans soutien adapté.

Nationality Group – Classification des Nationalités pour l'Analyse de la Réussite Académique

Basé sur un ensemble d'indicateurs de développement, d'éducation, de santé, de sécurité et de liberté économique, voici une classification exhaustive des nationalités pour analyser l'impact sur la réussite académique.

Groupe 0 : Pays de Très Haut Développement (Excellence Multi-Critères)

Scores Moyens : HDI > 0.95, Éducation > 0.93, Sécurité < 1.5, Corruption > 80, Liberté Économique > 75

01 : Élite Européenne Occidentale

- **2** : Allemand
- **13** : Néerlandais
- **14** : Anglais

Profil : HDI 0.95-0.96, systèmes éducatifs top 10 mondial, santé excellente, très faible corruption (80+), sécurité optimale. **Impact académique :** Réussite très élevée, accès privilégié aux ressources, environnement familial stimulant intellectuellement.

02 : Europe du Sud Développée

- **1** : Portugais
- **6** : Espagnol
- **11** : Italien

Profil : HDI 0.89-0.92, bons systèmes éducatifs, santé développée, corruption modérée (60-70), sécurité correcte. **Impact académique :** Bonne réussite, culture éducative familiale forte, mais ressources parfois limitées.

Groupe 1 : Pays de Haut Développement Émergent

Scores Moyens : HDI 0.80-0.94, Éducation 0.75-0.90, Sécurité 1.5-2.5, Corruption 40-70

10 : Europe de l'Est Intégrée

- **17** : Lituanien
- **62** : Roumain

Profil : HDI 0.85-0.89, éducation en amélioration, santé correcte, corruption en baisse (45-55), sécurité stable. **Impact académique :** Forte motivation éducative, valorisation de l'excellence, mais parfois contraintes financières.

11 : Puissances Émergentes

- **32** : Turc
- **105** : Russe

Profil : HDI 0.83-0.85, éducation inégale géographiquement, santé variable, corruption significative (30-45), sécurité préoccupante. **Impact académique :** Élites très performantes, mais disparités importantes selon milieu socio-économique.

Groupe 2 : Pays à Développement Intermédiaire

Scores Moyens : HDI 0.70-0.85, Éducation 0.60-0.80, Sécurité 2.0-3.0, Corruption 20-50

20 : Amérique Latine Émergente

- **41** : Brésilien
- **101** : Mexicain
- **109** : Colombien

Profil : HDI 0.75-0.79, éducation très inégalitaire, santé à deux vitesses, corruption élevée (25-35), violence préoccupante. **Impact académique :** Grande disparité entre classes sociales, élites performantes mais majorité défavorisée.

21 : Europe de l'Est en Transition

- **100** : Moldave
- **103** : Ukrainien

Profil : HDI 0.77-0.78, éducation héritée du système soviétique mais dégradée, santé en difficulté, corruption importante, contexte géopolitique instable. **Impact académique :** Solides bases académiques traditionnelles mais contexte socio-économique difficile.

30 : Systèmes Autoritaires Spécifiques

- **108** : Cubain

Profil : HDI 0.76, éducation universelle de qualité, santé remarquable malgré les moyens, corruption faible mais liberté limitée, sécurité correcte. **Impact académique :** Excellent niveau éducatif de base, forte culture scientifique, mais opportunités limitées.

Groupe 4 : Pays en Développement - Afrique Lusophone

Scores Moyens : HDI 0.45-0.67, Éducation 0.35-0.65, Sécurité 2.0-2.5, Corruption 15-35

40 : Îles Atlantiques Stabilisées

- **22** : Cap-Verdien
- **26** : Santoméen

Profil : HDI 0.66-0.64, petites économies insulaires, éducation en progrès, santé de base, corruption modérée, sécurité acceptable. **Impact académique :** Progrès éducatifs notables, valorisation de l'éducation mais ressources limitées.

41 : Géants Africains Ressources Naturelles

- **21** : Angolais
- **25** : Mozambicain

Profil : HDI 0.49-0.61, éducation très inégalitaire, santé défaillante, corruption très élevée (15-25), sécurité problématique post-conflit. **Impact académique :** Élites privilégiées mais majorité de la population avec accès limité à l'éducation de qualité.

42 : Petits États Fragiles

- **24** : Guinéen

Profil : HDI 0.50, système éducatif fragile, santé précaire, corruption endémique, instabilité politique récurrente. **Impact académique :** Défis majeurs, réussite dépendante du milieu familial urbain privilégié.

Facteurs Déterminants pour la Réussite Académique

Indicateurs Positifs Forts

1. **HDI > 0.90** : Corrélation +0.75 avec réussite universitaire
2. **Indice Éducation > 0.85** : Environnement familial stimulant
3. **Corruption < 40/100** : Mérite reconnu, système équitable
4. **Sécurité < 1.8** : Stabilité permettant les études longues

Facteurs de Compensation

1. **Élites éduquées** : Même dans pays Groupe 3-4, familles éduquées maintiennent performances
2. **Mobilité internationale** : Étudiants Groupes 2-4 souvent très motivés
3. **Réseaux diaspora** : Communautés établies facilitent adaptation
4. **Politiques éducatives** : Bourses et programmes spécifiques peuvent compenser origines

Profils de Risque Académique

Risque Faible (Groupes 1A-1B) : Réussite quasi-garantie, ressources abondantes **Risque Modéré (Groupe 2)** : Réussite probable avec soutien adapté **Risque Élevé (Groupe 3)** : Réussite variable, dépendante du milieu familial spécifique **Risque Très Élevé (Groupe 4)** : Réussite exceptionnelle, nécessite soutien intensif

Applications pour Prédiction d'Abandon

Cette classification multidimensionnelle permet d'identifier :

- Les étudiants nécessitant un **soutien financier** prioritaire (Groupes 3-4)
- Ceux bénéficiant d'**accompagnement linguistique/culturel** (Groupes 2B-4)
- Les **programmes de mentorat** ciblés selon l'origine
- Les **stratégies de rétention** adaptées aux profils de risque

---

Mother's qualification Group / Father's qualification Group – Classification Académique des Niveaux d'Éducation Parentale

Groupe 0 : Diplômes universitaires supérieurs

- 5 : Doctorat (3e cycle)
- 44 : Enseignement supérieur - doctorat (3e cycle)
- 4 : Master (2e cycle)
- 43 : Enseignement supérieur - master (2e cycle)
- 3 : Diplôme supérieur deuxième cycle
- 41 : Cours d'études supérieures spécialisées
- 40 : Enseignement supérieur - diplôme (1er cycle)
- 2 : Diplôme supérieur premier cycle

Caractéristiques : Parents très diplômés, souvent cadres, enseignants ou professions intellectuelles supérieures. Leurs enfants ont les meilleures chances d’accéder à l’enseignement supérieur et d’obtenir des diplômes élevés.

Groupe 1 : Diplômes secondaires et techniques

- 1 : Enseignement secondaire (12e année ou équivalent)
- 9 : 12e année non complétée
- 10 : 11e année non complétée
- 12 : Autres formations 11e année
- 13 : 2e année complémentaire lycée
- 14 : 10e année
- 20 : Formation complémentaire lycée
- 22 : Formation technique professionnelle
- 27 : 2e cycle du lycée général
- 39 : Cursus de spécialisation technologique
- 42 : Cours technique supérieur professionnel

Caractéristiques : Parents ayant suivi un cursus secondaire ou technique, professions intermédiaires ou techniciens. Les enfants ont des chances moyennes à élevées d'accéder à l'enseignement supérieur, mais moins aux diplômes les plus élevés.

Groupe 2 : Éducation de base et formations complémentaires

- 19 : Éducation de base 3e cycle
- 25 : Formation complémentaire non complétée
- 26 : 7e année
- 29 : 9e année non complétée
- 30 : 8e année
- 31 : Cours généraux administration et commerce
- 33 : Comptabilité et administration
- 37 : Éducation de base 1er cycle
- 38 : Éducation de base 2e cycle
- 18 : Commerce général
- 6 : Formation continue

Caractéristiques : Parents ayant une éducation de base ou des formations complémentaires courtes. Les enfants ont des chances plus faibles d'accéder à l'enseignement supérieur, mais peuvent réussir avec un fort soutien familial.

Groupe 3 : Faible niveau d'études ou inconnu

- 11 : 7e année (ancienne échelle)
- 34 : Inconnu
- 35 : Analphabète
- 36 : Lecture sans formation complète

Caractéristiques : Parents peu ou pas diplômés, souvent employés ou ouvriers. Les enfants sont sous-représentés dans l'enseignement supérieur et ont un risque plus élevé d'abandon scolaire.

Mother's occupation Group / Father's occupation Group – Classification des Professions Parentales selon l'Exigence Cognitive et l'Impact sur la Réussite Scolaire

Basé sur les recherches scientifiques concernant les corrélations entre professions parentales, niveau socio-économique et réussite académique, voici une classification des métiers de votre dataset selon leurs exigences cognitives et leur association avec la réussite scolaire des enfants.

Groupe 0 : Exigences Cognitives Très Élevées

Caractéristiques : formation universitaire supérieure, travail analytique complexe

- **1** : Représentants du pouvoir législatif et des organes exécutifs, directeurs et cadres dirigeants
- **2** : Spécialistes des activités intellectuelles et scientifiques
- **101** : Officiers des forces armées

- **121** : Spécialistes des sciences physiques, des mathématiques, de l'ingénierie et des techniques connexes
- **122** : Professionnels de la santé
- **123** : Enseignants
- **124** : Spécialistes en finance, comptabilité, organisation administrative, relations publiques et commerciales

Impact sur la réussite académique des enfants : Les études montrent que les enfants de ces professions ont 40-60% plus de chances de réussir dans l'enseignement supérieur. Ces parents offrent un environnement intellectuel stimulant, des ressources éducatives et des attentes académiques élevées.

Groupe 1 : Exigences Cognitives Élevées

Caractéristiques : formation technique supérieure ou universitaire, résolution de problèmes complexes

- **3** : Techniciens et professions de niveau intermédiaire
- **102** : Sous-officiers des forces armées
- **112** : Directeurs des services administratifs et commerciaux
- **114** : Directeurs de l'hôtellerie, de la restauration, du commerce et d'autres services
- **131** : Techniciens et professions intermédiaires des sciences et de l'ingénierie
- **132** : Techniciens et professionnels de santé de niveau intermédiaire
- **134** : Techniciens de niveau intermédiaire des services juridiques, sociaux, sportifs, culturels et similaires
- **135** : Techniciens en technologies de l'information et de la communication

Impact sur la réussite académique des enfants : Ces familles montrent des taux de réussite académique supérieurs à la moyenne (25-35% d'amélioration). L'éducation parentale compense souvent les revenus plus modestes par rapport au Groupe 1.

Groupe 2 : Exigences Cognitives Moyennes-Élevées

Caractéristiques : formation professionnelle spécialisée, compétences techniques

- **4** : Personnel administratif
- **103** : Autre personnel des forces armées
- **141** : Employés de bureau, secrétaires et opérateurs de saisie de données
- **143** : Opérateurs de services de données, de comptabilité, statistiques, financiers et de registre
- **144** : Autre personnel de soutien administratif
- **154** : Personnel des services de protection et de sécurité
- **174** : Ouvriers qualifiés en électricité et en électronique

Impact sur la réussite académique des enfants : Impact modéré mais positif sur la réussite scolaire. Ces professions offrent stabilité et structure, favorisant la discipline d'étude chez les enfants.

Groupe 3 : Exigences Cognitives Moyennes

Caractéristiques : formation professionnelle, compétences pratiques et relationnelles

- **5** : Travailleurs des services personnels, de la sécurité et des ventes
- **151** : Travailleurs des services personnels
- **152** : Vendeurs

- **153** : Travailleurs des soins personnels et assimilés
- **161** : Agriculteurs de culture de marché et ouvriers qualifiés de la production agricole et animale
- **171** : Ouvriers qualifiés du bâtiment et assimilés, à l'exception des électriciens
- **172** : Ouvriers qualifiés de la métallurgie, du travail des métaux et similaires
- **175** : Travailleurs de la transformation alimentaire, du bois, de l'habillement et autres industries et artisanats
- **181** : Opérateurs d'installations fixes et de machines
- **182** : Travailleurs de l'assemblage
- **183** : Conducteurs de véhicules et opérateurs d'équipements mobiles

Impact sur la réussite académique des enfants : Impact neutre à légèrement positif selon l'environnement familial. La motivation et les valeurs familiales jouent un rôle plus important que le niveau professionnel.

Groupe 4 : Exigences Cognitives Moyennes-Faibles

Caractéristiques : formation de base, travail routinier

- **6** : Agriculteurs et ouvriers qualifiés de l'agriculture, de la pêche et de la sylviculture
- **7** : Ouvriers qualifiés de l'industrie, de la construction et artisans
- **8** : Opérateurs d'installation et de machines et travailleurs de l'assemblage
- **163** : Agriculteurs, éleveurs, pêcheurs, chasseurs et cueilleurs de subsistance
- **192** : Travailleurs non qualifiés de l'agriculture, de la production animale, de la pêche et de la sylviculture
- **193** : Travailleurs non qualifiés de l'industrie extractive, de la construction, de la fabrication et du transport
- **194** : Aides à la préparation des repas
- **195** : Vendeurs ambulants (hors produits alimentaires) et prestataires de services de rue

Impact sur la réussite académique des enfants : Les recherches montrent des défis mais pas une impossibilité de réussite. Avec un soutien scolaire approprié, les enfants peuvent surmonter les désavantages socio-économiques.

Groupe 5 : Exigences Cognitives Faibles

Caractéristiques : travail principalement manuel et routinier

- **9** : Travailleurs non qualifiés

Impact sur la réussite académique des enfants : Défis significatifs mais surmontables avec intervention précoce et soutien éducatif intensif. Les études montrent l'importance cruciale de l'école dans ces cas.

Groupe 6 : Groupes Non Classifiables par Exigences Cognitives

- **0** : Étudiant (en formation)
- **10** : Professions des forces armées (hiérarchie variable)
- **90** : Autre situation
- **99** : (vide)

Curricular units 1st sem (Approved/Enrolled) et Curricular units 2nd sem (Approved/Enrolled) – Taux de Réussite Académique par Semestre

Le taux de réussite académique par semestre est un indicateur clé pour évaluer la performance des étudiants dans leurs unités d'enseignement respectives. Il est directement lié au nombre d'unités validées par rapport au nombre total d'unités auxquelles l'étudiant s'est inscrit ce qui renvoit à son efficacité académique.

Calcul du Taux de Réussite par Semestre :

- Taux de Réussite 1er Semestre = (Curricular units 1st sem (approved) / Curricular units 1st sem (enrolled)) * 100
- Taux de Réussite 2nd Semestre = (Curricular units 2nd sem (approved) / Curricular units 2nd sem (enrolled)) * 100

---

4. Features Supprimés :

La plupart des variables initiales ont été conservées. Cependant, les variables suivantes ont été supprimées en raison de leur remplacement par des features de groupe plus informatives :

- Previous qualification
- Nacionality
- Mother's qualification
- Father's qualification
- Mother's occupation
- Father's occupation
- Course (171)
    - Supprimé car aucune évaluation n'a été réalisée pour cette filière spécifique dans le dataset.

---

5. Caractéristiques des données après nettoyage et pré-traitement

- **Nombre total de lignes après préparation** : 4181
- **Nombre total de colonnes après préparation** : 39
- **Valeurs manquantes** : Aucune
- **Anomalies détectées** :
    - Suppression des lignes où "Mother's occupation" = [125, 173, 191] --> aucune correspondance métier n'est trouvée.
    - Suppression des lignes où "Father's occupation" = [125, 173, 191] --> aucune correspondance métier n'est trouvée.
    - **Nombre / Pourcentage de lignes supprimées** : 28 / 0.63 %
    - Suppression des lignes où course = [171] --> aucune évaluation réalisée.
    - **Nombre / Pourcentage de lignes supprimées** : 215 / 4.89 %
- **Répartition des classes après séparation** :
    - Graduate : 2097 (50.2 %) --> +0.2 par rapport aux données initiales (2209)
    - Dropout : 1339 (32.0 %) --> -0.1 par rapport aux données initiales (1421)
    - Enrolled : 745 (17.8 %) --> -0.1 par rapport aux données initiales (794)

---

6. Séparation en deux fichiers distincts - Enrolled vs Graduate/Dropout

Nous devons séparer les fichiers en deux ensembles distincts pour différentes analyses car les étudiants "Enrolled" (toujours inscrits) ne sont pas pertinents pour l'analyse des taux de réussite ou d'abandon. Ils pourront nous être utiles pour des analyses futures sur la prédiction de la réussite académique en cours d'études.

Après analyses des données, 2 fichiers CSV ont été crées :

- `data/data_enrolled.csv` : Contient uniquement les étudiants inscrits (Enrolled).
  - Nombre d'observations : 782
  - Répartition des cibles :
    - Enrolled : 745 (100 %)
- `data/data_graduate_dropout.csv` : Contient les étudiants diplômés (Graduate) et ceux ayant abandonné (Dropout).
  - Nombre d'observations : 3436
  - Répartition des cibles :
    - Graduate : 2097 (61 %)
    - Dropout : 1339 (39 %)

---

7. Séparation en train et test

Pour le fichier `data/data_graduate_dropout.csv`, une séparation en ensembles d'entraînement et de test a été effectuée pour les analyses prédictives futures.

- Taille de l'ensemble d'entraînement : 80 % (observations)
- Taille de l'ensemble de test : 20 % (observations)

---

8. Encodage et Standardisation des features

En raison de la présence de variables catégorielles dans le dataset, un encodage one-hot sans drop first a été appliqué pour convertir ces variables en un format numérique adapté aux algorithmes de machine learning.

Application de l'encodage OneHot sur les variables suivantes :

- Marital Status
- Application mode
- Course
- Daytime/evening attendance
- Displaced
- Educational special needs
- Debtor
- Tuition fees up to date
- Gender
- Scholarship holder
- International

Application du label encoding sur les variables suivantes :

- target

Application de l'encodage ordinal sur les variables suivantes :

- Application order
- Previous qualification Group
- Mother's qualification Group
- Father's qualification Group
- Admission grade
- Age at enrollment
- Curricular units 1st sem (credited)
- Curricular units 1st sem (enrolled)", "Curricular units 1st sem (evaluations)",
- Curricular units 1st sem (evaluations)
- Curricular units 1st sem (approved)
- Curricular units 1st sem (grade)
- Curricular units 1st sem (without evaluations)
- Curricular units 2nd sem (credited)
- Curricular units 2nd sem (enrolled)
- Curricular units 2nd sem (evaluations)
- Curricular units 2nd sem (approved)
- Curricular units 2nd sem (grade)
- Curricular units 2nd sem (without evaluations)
- Unemployment rate
- Inflation rate
- GDP

Nombre de features après encodage (OneHot encoding) 83 features + 1 target

---

9. Standardisation des données

Pour assurer que toutes les features contribuent de manière égale aux analyses prédictives, une standardisation des données a été effectuée. Chaque feature a été transformée pour avoir une moyenne de 0 et un écart-type de 1. Cette étape est cruciale pour les algorithmes sensibles à l'échelle des données, tels que la régression logistique et les réseaux de neurones.

---

10. Références

Informations sur les auteurs et financement

- **Créé dans le cadre** : Programme SATDAP - Capacitação da Administração Pública sous la subvention POCI-05-5762-FSE-000191, Portugal.

- **Auteurs principaux** :
    - Mónica Vieira Martins (Instituto Politécnico de Portalegre)
    - Daniel Tolledo (Instituto Politécnico de Portalegre)
    - Jorge Machado (Instituto Politécnico de Portalegre)
    - Luís M. T. Baptista (Instituto Politécnico de Portalegre)

- Valentim Realinho (Instituto Politécnico de Portalegre - VALORIZA - Research Center for Endogenous Resource Valorization, Portalegre, Portugal) **Citation complète**: Martins, M.V., Tolledo, D., Machado, J., Baptista, L.M.T., Realinho, V. (2021). Early Prediction of student's Performance in Higher Education: A Case Study. In: Rocha, Á., Adeli, H., Dzemyda, G., Moreira, F., Ramalho Correia, A.M. (eds) Trends and Applications in Information Systems and Technologies. WorldCIST 2021. Advances in Intelligent Systems and Computing, vol 1365. Springer, Cham. https://doi.org/10.1007/978-3-030-72657-7_16 Lien : text

---

11. Licence