深度剖析「預測學生輟學與學業成功」資料集 (UCI 資料集 697)

1.「預測學生輟學與學業成功」資料集簡介

「預測學生輟學與學業成功」資料集，登錄於 UCI 機器學習儲存庫 (UCI Machine Learning Repository)，編號為 697，為研究高等教育中學生續存及學業成果相關因素提供了豐富的資訊來源 [1]。此資料集於 2021 年 12 月 12 日捐贈 [3]。其核心內容源自一所高等教育機構，旨在建立分類模型以預測學生輟學與學業成功 [1]。資料集涵蓋了學生入學時的已知資訊（包括學術背景、人口統計學資料以及社會經濟因素）以及他們在第一和第二學期結束時的學業表現 [2]。此類資料集的出現，對於教育資料探勘 (Educational Data Mining, EDM) 與學習分析 (Learning Analytics, LA) 領域至關重要，使研究人員能夠運用計算方法來理解並改進教育過程。此資料集的現實意義在於其直接回應了全球高等教育領域中學生輟學與學業失敗這一嚴峻議題。這些問題對個人、教育機構乃至整個社會都具有深遠的影響 [3]。資料集的創建目標明確指出其旨在減少學術輟學與失敗 [3]。多方資料均描述輟學為全球性問題，不僅是經濟增長的障礙，亦是造成社會困境的原因，並對教育機構的聲譽與運作成效構成挑戰 [5]。透過理解與預測這些學業成果，可以為高風險學生制定及時的介入措施與支持機制。該資料集的存在及其用途，突顯了教育實踐上的一項重大轉變：從以往對學生失敗的被動反應，轉向更為主動、以數據驅動的干預措施。資料集明確設計用於「在學術路徑的早期階段識別高風險學生」[3]，這本身即體現了一種前瞻性的策略。基於此資料集開發的「學習分析工具」[6] 進一步鞏固了此趨勢。這種主動姿態與傳統教育模式（可能等待失敗發生後才介入）形成對比，其更廣泛的意涵是朝向個性化教育與支持發展，利用數據預期學生需求。然而，儘管此資料集旨在支持學生，其在預測模型中的應用本身也引發了關於標籤化、偏見以及若干預措施設計與實施不當可能導致自我實現預言等倫理問題。其目標是「識別高風險學生」[3]，這種識別雖然有助於針對性支持，但若處理不當，也可能導致污名化或差別待遇。資料集包含人口統計學與社會經濟因素 [1]。若基於這些數據訓練的模型延續了與這些因素相關的現有偏見，則可能無意中傷害了它們旨在幫助的學生。在考慮應用由此資料集建構的模型時，這是一個關鍵的後續討論領域。

2. 資料集的起源、目標與重要性

2.1 創建背景 此資料集源於葡萄牙一所高等教育機構（波塔萊格雷理工學院，Polytechnic Institute of Portalegre）的一個研究計畫 [1]。該計畫旨在運用機器學習技術減少學術輟學現象 [1]。更具體地，如 [6] 所述，該機構開發了一套學習分析工具，此資料集即為該工具的一部分。[34] 亦確認了該機構以及資料收集的時間範圍為 2008 年至 2018 年。**2.2** 資金來源 該計畫獲得了葡萄牙「公共行政能力建設計畫」(SATDAP - Capacitação da Administração Pública) 的支持，撥款編號為 POCI-05-5762-FSE-000191 [1]。註明資金來源是學術研究的標準做法，同時也突顯了政府或機構對此類研究計畫的支持。公共資金的投入通常伴隨著特定的任務或期望成果。「公共行政能力建設計畫」的名稱本身即暗示了其重點在於改善公共服務，在此案例中即為高等教育。這表明資料集的創建是應對公認國家問題（學生流失）的策略性措施的一部分，將研究直接與政策目標聯繫起來。**2.3** 設定目標 資料集的主要目標是建立分類模型，以預測學生的輟學情況與學業成功 [1]。此預測問題被設定為一個三分類任務：「輟學 (dropout)」、「在學 (enrolled)」和「畢業 (graduate)」，評估時間點為課程正常修業年限結束時 [1]。一個核心宗旨在於早期識別高風險學生，以便及時實施支持策略 [1]。儘管「早期識別」（如眾多資料來源所述）是關鍵目標，但資料集主要包含入學及第一、二學期的數據。挑戰在於預測模型能否基於這些相對早期的數據，準確預測長期結果（3-4 年後的輟學/畢業）。「早期性」是一把雙面刃：有利於及時干預，但由於學生學習旅程中不可預見的

變化，其準確性可能較低。資料集的結構本身側重於早期可知的信息，這在實務上是可行的，但可能限制了對更長時期內所有影響因素的全面理解。[47]、[26]、[26] 及 [40] 的研究探討了預測與干預的最佳時機，發現在早期階段中，第一學期末的數據是最佳的。**2.4** 重要性與意義 此資料集有助於減少學術輟學與學業失敗，這些是重大的社會與經濟問題 [1]。它使得針對高風險學生開發目標明確的干預措施與支持策略成為可能 [1]。此外，它亦可作為比較機器學習演算法效能的基準資料集，以及教育資料探勘/學習分析領域的訓練材料 [6]。波塔萊格雷理工學院利用此資料集開發「學習分析工具」[6]，顯示該機構除了為一般研究做出貢獻外，亦有強烈的內部動機去理解其學生群體並改進其支持系統。為其「輔導團隊」[6] 創建特定工具，表明資料集在機構內部有直接應用，用於加強學生支持和提升機構效能。這突顯了一種務實的、以應用為導向的研究焦點。

3. 整體資料概覽

3.1 實例數量 資料集包含 4424 筆記錄，每筆記錄代表一名獨立的學生 [3]。多個來源均確認此數量，例如 [22]、[3]、[6] 及 [7]。[3] 及 [3] 亦明確指出「每個實例都是一名學生」。**3.2** 屬性數量 關於屬性數量，不同來源的資訊略有差異。UCI 頁面及部分衍生資料提及 36 個特徵 [36] 或 36 個屬性 [5]。而 MDPI 的論文 [6] 及一些 Kaggle 描述 [22] 則指出有 35 個屬性。這種差異可能源於目標變數是否被計為屬性或特徵。UCI 頁面 [3] 將「目標 (Target)」列為與「特徵 (features)」分開的組件。MDPI 論文 [6] 列出了 34 個特徵加上 1 個目標變數，總共 35 個。一個常見的 Kaggle 版本 (dataset.csv) 有 35 個欄位 [28]，包括目標變數。Martins 等人 (2021) 的論文（引自 [34]）提及 37 個特徵，這可能包含一個 ID 或是一個略有不同的版本。為求清晰，本報告將採納 34 個輸入特徵和 1 個目標變數的結構，此與 MDPI 論文 [6] 的詳細內容及 UCI 主頁結構 [3] 一致。[24] 提及 36 個屬性，後又稱「36 個自變數和一個因變數」，這造成了混淆。最一致且詳細的來源 [6] 指向 34 個特徵 + 1 個目標變數。這些在不同來源中（35、36、37 個）[3] 報導的屬性數量略有差異，突顯了資料集文檔記錄與版本控制中常見的挑戰。雖然差異微小，但研究人員必須仔細交叉引用。這意味著使用者必須警惕他們正在使用的資料集特定版本，並依賴最詳細的附帶文檔（如 MDPI 論文或 UCI 變數表）來獲取權威的屬性列表。這也暗示如果研究人員不夠謹慎，不同的研究論文可能正在使用略有不同的特徵集。**3.3** 資料集類型 此資料集屬於多變量 (Multivariate) 類型 [3]。[61] 雖然指的是另一個學生資料集，但也提供了此分類的範例。**3.4** 遺失值 最終的資料集明確指出不包含任何遺失值 [3]。創作者進行了嚴謹的預處理以處理遺失值 [3]。明確聲明「已執行嚴謹的資料預處理以處理異常值、無法解釋的離群值和遺失值」[3]，這顯著提升了資料集直接用於機器學習模型的品質與可靠性，減輕了研究人員初步資料清理的負擔。資料集通常需要大量清理工作。創作者致力於提供一個乾淨的資料集（無遺失值 [3]），意味著研究人員可以更專注於特徵工程、模型選擇和分析，而非基本的資料插補或清理。這使得資料集更易於取用，且不同研究之間的結果可能更具可比較性，前提是預處理過程是健全的。**3.5** 代表性與普遍性 此資料集包含來自葡萄牙單一理工學院的 4424 個實例 [6]。雖然具有價值，但基於此數據訓練的模型的普遍性引發了疑問，即這些模型是否能推廣到其他具有不同人口統計學、社會經濟和學術背景的機構、國家或教育系統。教育系統、學生群體和影響因素可能差異很大。因此，使用此資料集開發的模型在其他地方可能表現不佳，或需要大量重新校準。這是教育資料探勘研究中常見的限制，但承認這一點很重要。[26] 和 [26] 提到目標是「推廣到該理工大學的任何學位」，這是一個良好的開端，但要超越該機構進行推廣則是一個更大的挑戰。

4. 深入屬性分析

本節將呈現所有屬性（特徵與目標變數）的完整列表。此列表的主要來源是 Mendes 提供的 PDF 文件中的詳細資料字典 [50]，該文件似乎是研究材料中關於編碼最詳盡的列表。這與

UCI 頁面 [3] 及 MDPI 論文 [6] 的資訊高度吻合。這些屬性通常被分為：人口統計資料、社會經濟資料、宏觀經濟資料、入學時學術資料、第一學期末學術資料、第二學期末學術資料以及目標變數 [6]。

表 1：詳細屬性字典

本表旨在提供對每個變數及其編碼的清晰、明確的理解，這對於任何資料分析或機器學習任務都至關重要。若無此理解，詮釋可能出現偏差，模型輸入也可能不正確。資料集對許多類別型特徵（例如婚姻狀況、申請模式、課程）使用數值代碼。本表將這些代碼轉換為人類可讀的意義，這對於探索性資料分析 (EDA)、特徵工程以及詮釋模型結果（例如特徵重要性）至關重要。它將分散在不同來源（UCI 頁面、MDPI 論文、其他文件如 [50]）的資訊集中到單一、全面的參考資料中。這份詳細的字典使研究人員能夠準確地預處理資料（例如，決定使用獨熱編碼還是序數編碼）、選擇相關特徵，並理解其研究結果的背景。例如，了解「性別：1 – 男性；0 – 女性」[50] 對於詮釋任何與性別相關的模型輸出至關重要。這直接解決了此資料集使用者常見的困惑點，Kaggle 討論區中即有使用者詢問數值標籤的意義 [51]。

變數名稱 (Variable Name)	角色 (Role)	類型 (Type)	描述 (Description)	編碼/特定值 (Encoding / Specific Values)
Marital status	Feature	Categorical (Nominal)	學生的婚姻狀況	1=未婚 (single), 2=已婚 (married), 3=鰥寡 (widower), 4=離婚 (divorced), 5=事實婚姻 (facto union), 6=合法分居 (legally separated) [3]
Application mode	Feature	Categorical (Nominal)	學生的申請方式	數值編碼，例如：1=第一階段-普通名額 (1st phase - general contingent), 2=第612/93號條例 (Ordinance No. 612/93),..., 57=機構/課程變更(國際生) (Change of institution/course (International)) [3]
Application order	Feature	Numerical (Ordinal)	學生申請的順序	0=第一志願 (first choice) 至 9=最後志願 (last

				choice) [3]。實際資料集中的值域應進行驗證，因部分來源僅列為數值型。
Course	Feature	Categorical (Nominal)	學生修讀的課程	數值編碼，例如：33=生物燃料生產技術 (Biofuel Production Technologies), 171=動畫與多媒體設計 (Animation and Multimedia Design),..., 9991=管理學(夜間) (Management (evening attendance)) [3]
Daytime/evening attendance	Feature	Binary	學生是日間部或夜間部	1=日間 (daytime), 0=夜間 (evening) [3]
Previous qualification	Feature	Categorical (Ordinal)	學生入學前的學歷	數值編碼，例如：1=中學教育 (Secondary education), 2=高等教育-學士學位 (Higher education - bachelor's degree),..., 43=高等教育-碩士(第二週期) (Higher education - master (2nd cycle)) [3]
Previous qualification	Feature	Numerical (Continuous)	入學前學歷的成績	0 至 200 之間 [3]

(grade)				
Nacionality	Feature	Categorical (Nominal)	學生的國籍	數值編碼，例如：1=葡萄牙籍 (Portuguese), 2=德國籍 (German),..., 109=哥倫比亞籍 (Colombian) [3]
Mother's qualification	Feature	Categorical (Ordinal)	學生母親的學歷	數值編碼，代表教育程度 [503]
Father's qualification	Feature	Categorical (Ordinal)	學生父親的學歷	數值編碼，代表教育程度 [503]
Mother's occupation	Feature	Categorical (Nominal)	學生母親的職業	數值編碼 [503]
Father's occupation	Feature	Categorical (Nominal)	學生父親的職業	數值編碼 [503]
Admission grade	Feature	Numerical (Continuous)	入學成績，用於評估申請者整體學術適用性	0 至 200 之間 [50]
Displaced	Feature	Binary	學生是否為流離失所者	1=是 (yes), 0=否 (no) [3]
Educational special needs	Feature	Binary	學生是否有特殊教育需求	1=是 (yes), 0=否 (no) [3]
Debtor	Feature	Binary	學生是否為債務人	1=是 (yes), 0=否 (no) [3]
Tuition fees up to date	Feature	Binary	學生的學費是否繳清	1=是 (yes), 0=否 (no) [3]
Gender	Feature	Binary	學生的性別	1=男性 (male), 0=女性 (female) [3]

Scholarship holder	Feature	Binary	學生是否持有獎學金	1=是 (yes), 0=否 (no) [3]
Age at enrollment	Feature	Numerical (Discrete)	學生入學時的年齡	[3]
International	Feature	Binary	學生是否為國際學生	1=是 (yes), 0=否 (no) [3]
Curricular units 1st sem (credited)	Feature	Numerical (Discrete)	第一學期學生獲得的學分數	[3]
Curricular units 1st sem (enrolled)	Feature	Numerical (Discrete)	第一學期學生註冊的學分數	[3]
Curricular units 1st sem (evaluations)	Feature	Numerical (Discrete)	第一學期學生接受評估的課程單元數	[3]
Curricular units 1st sem (approved)	Feature	Numerical (Discrete)	第一學期學生通過的課程單元數	[3]
Curricular units 1st sem (grade)	Feature	Numerical (Continuous)	第一學期課程單元的平均成績	0 至 200 之間 [3]
Curricular units 1st sem (without evaluations)	Feature	Numerical (Discrete)	第一學期未進行評估的課程單元數	[3]
Curricular units 2nd sem (credited)	Feature	Numerical (Discrete)	第二學期學生獲得的學分數	[3]
Curricular units 2nd sem (enrolled)	Feature	Numerical (Discrete)	第二學期學生註冊的學分數	[3]
Curricular units 2nd sem (evaluations)	Feature	Numerical (Discrete)	第二學期學生接受評估的課程單元數	[3]
Curricular units 2nd sem	Feature	Numerical (Discrete)	第二學期學生通過的課程單元數	[3]

(approved)				
Curricular units 2nd sem (grade)	Feature	Numerical (Continuous)	第二學期課程單元的平均成績	0 至 200 之間 [3]
Curricular units 2nd sem (without evaluations)	Feature	Numerical (Discrete)	第二學期未進行評估的課程單元數	[3]
Unemployment rate	Feature	Numerical (Continuous)	該地區的失業率	[6]
Inflation rate	Feature	Numerical (Continuous)	該地區的通貨膨脹率	[6]
GDP	Feature	Numerical (Continuous)	該地區的國內生產總值	[6]
Target	Target	Categorical (Nominal)	學生在正常修業年限結束時的狀態	Dropout (輟學), Enrolled (在學), Graduate (畢業) [1]

資料集為第一和第二學期提供了非常詳細的學業表現細目（已獲學分、已註冊學分、評估課程數、通過課程數、成績、未評估課程數）。這種細緻程度是一大優勢，有助於對早期學術參與和成功/困境的不同方面如何影響長期成果進行細緻分析 [3]。研究人員可以藉此探討諸如：通過課程的「數量」是否比「平均成績」更具預測性？高註冊學分數伴隨低通過學分數是否為風險信號？這些詳細數據支持更複雜的特徵工程和對學術進展的更深入理解。這或許解釋了為何「第二學期通過課程數 (Curricular units 2nd sem (approved))」和「第一學期通過課程數 (Curricular units 1st sem (approved))」等變數在預測模型中常被視為高度重要的特徵 [13]。資料集納入了父母學歷/職業、獎學金狀況、債務狀況、學費繳納狀況以及地區宏觀經濟指標（失業率、通貨膨脹率、GDP）[3]。這明確承認學生的成功不僅取決於個人學術能力，也深植於更廣泛的社會經濟背景之中。諸如「學費是否繳清 (Tuition fees up to date)」[3] 等因素直接將財務狀況與續讀意願聯繫起來。父母學歷（「母親學歷 (Mother's qualification)」、「父親學歷 (Father's qualification)」[3]）可作為社會經濟地位和與教育相關的文化資本的代理變數。宏觀經濟因素 [6] 可能影響就業前景（構成輟學的拉力因素 [24]）或整體財務壓力。這種多層次數據有助於探索個人、家庭和社會因素之間的交互作用。「入學前學歷 (Previous qualification)」屬性包含大量類別 [3]，代表了多樣化的教育背景。這種豐富性在建模時可能難以處理（需要仔細編碼或分組），但也提供了研究從各種教育流向進入高等教育的過渡路徑的機會。其編碼 [3] 涵蓋了從「中學教育」到「技術專業課程」再到「高等教育 - 碩士」（可能適用於攻讀另一學位的學生）等。這種多樣性意味著簡單地將其視為單一序數變數可能過於簡化。研究人員可能需要有意義地對這些類別進行分組，或使用更複雜的編碼方案來捕捉先前教育路徑的真實影響。這也引發了關於不同先前學歷如何為學生在該機構特定課程中的成功做好準備的問題。[19] 和 [19] 指出，在某些模型中，先前學歷的預測能力較低，這可能是由於其複雜性，或是因為當前的學業表現掩蓋了它的影響。

5. 資料來源與預處理

5.1 原始資料來源 此資料集是從數個互不相干的資料來源彙整而成[1]。這些來源包括：

- 波塔萊格雷理工學院的內部系統：
  - 學術管理系統 (Academic Management System, AMS)[6]。
  - 教學活動支援系統 (Support System for the Teaching Activity, PAE - 內部開發)[6]。
- 外部資料：
  - 高等教育總署 (General Directorate of Higher Education, DGES) - 關於透過全國高等教育入學競賽 (National Competition for Access to Higher Education, CNAES) 的入學資料[6]。
  - 葡萄牙當代資料庫 (Contemporary Portugal Database, PORDATA) - 提供宏觀經濟數據[6]。這種多來源資料彙整（AMS、PAE、DGES、PORDATA）[6] 突顯了創建全面學生資料集所需付出的巨大努力。教育資料通常分散在孤立的系統中，整合這些資料在技術和後勤上都是一項不小的挑戰。每個來源（AMS、PAE、DGES、PORDATA）很可能有不同的格式、識別碼和更新週期。創建自訂 VBA 程式以及特定的資料清理步驟（例如，處理來自 AMS 的 13,992 列和 398 行數據）[6] 更突顯了其複雜性。這意味著此類資料集的製作和維護成本高昂，使得像這樣公開可用的資料集尤為珍貴。

5.2 資料收集方法 資料涉及 2008/2009 學年至 2018/2019 學年期間入學的學生[6]。此時期意義重大，因為它是在歐洲高等教育實施博洛尼亞進程 (Bologna Process) 之後[7]。資料收集時間點在「歐洲高等教育實施博洛尼亞進程之後」[7] 這一點意義重大。博洛尼亞進程旨在標準化歐洲高等教育。此時間點表明資料集反映了一種更為協調的教育結構，與博洛尼亞改革前的數據相比，可能使得研究結果在其他同樣採納博洛尼亞改革的歐洲機構之間更具可比較性。博洛尼亞進程引入了諸如三週期制（學士、碩士、博士）等變革。此時代的數據將反映這些結構性要素。對於希望比較研究結果或在不同歐洲背景下應用模型的研究人員而言，此背景非常重要，它提供了一個共同的結構基準。資料集涵蓋來自不同知識領域的 17 個大學學位課程的學生，例如農學、設計、教育、護理、新聞、管理、社會服務和技術等[2]。**5.3** 資料集創建者的特定預處理步驟 根據[6]、[6]、[7]及[7]的詳細說明，預處理步驟如下：

- 準備全國入學競賽資料 **(CNAES)**：開發了一個 VBA 程式，從 DGES 每年提供的 Microsoft Access 資料庫中收集資訊，並匯出為 competition.csv 檔案。此檔案涵蓋了「入學時資料」屬性。
- 準備學生記錄資料 **(AMS)**：
  - 從 AMS 收到的初始 CSV 檔案包含 13,992 列和 398 行，其中有許多重複或不相關的條目。
  - 刪除了已停止招生之舊課程的學生記錄，以及不相關入學方式（例如 Erasmus 交換生）的學生記錄。
  - 選擇並重新命名了相關欄位，並移除了重複列。

    - 此步驟收集了「人口統計資料」和「社會經濟資料」。
- 準備學生評估資料 **(PAE)**：處理了包含學生評估嘗試資訊的 CSV 檔案，以計算與「第一學期末學術資料」和「第二學期末學術資料」相關的屬性。
- 合併與預處理資料：
    - 合併了前述步驟中收集的所有資料。
    - 加入了宏觀經濟資料（來自 PORDATA）。
    - 執行了嚴謹的資料預處理，以處理異常值、無法解釋的離群值和遺失值 [3]。這是確保資料品質的關鍵步驟。
    - 最後，將學生分類至目標變數（輟學、在學、畢業）。
- 最終資料集為 UTF8 編碼的 CSV 檔案 [6]。[24] 提及已修改為 CSV 格式。

特定的預處理決策（例如，「刪除已註冊舊課程且目前不再招生的學生記錄，刪除入學方式不相關的學生記錄，如 Erasmus 交換生」）[6] 塑造了最終的學生群體，وبالتالي，也影響了基於此資料集訓練的模型。例如，移除 Erasmus 交換生使資料集聚焦於該機構的常規、攻讀學位的學生。移除舊課程確保了與當前課程的相關性。這些選擇對於既定目標而言是合理的，但也意味著資料集並不能代表所有曾與該機構互動過的學生，而僅代表一個特定的、相關的子集。這影響了可以得出的結論範圍。處理「異常值、無法解釋的離群值和遺失值」[3] 至關重要；雖然它清理了數據，但所使用的方法（如果未完全詳細說明）可能會引入偏見或移除真實存在但極端的有效案例。

6. 定義學生結果：目標變數

6.1 變數名稱 目標變數的名稱為 Target [1]。**6.2** 性質 此變數為類別型，具有三個類別 [1]。**6.3** 類別定義 根據 [6]、[7] 及 [7]，類別定義採微觀視角 (micro-perspective)，評估時間點為課程正常修業年限結束時（一般為 3 年，護理學程為 4 年）：

- 輟學 **(Dropout)**：指未完成學業計畫、未取得畢業學位即離校的學生。無論何時發生，轉換領域或學校均視為輟學。此定義較宏觀視角 (macro-perspective) 更為寬泛 [6]。相較於僅考慮完全離開高等教育體系而未取得學位的學生之宏觀定義，此微觀定義將導致更高的輟學率。此選擇直接影響「輟學」類別的性質以及模型訓練的預測目標。這對於詮釋模型結果以及設計干預措施（例如，旨在讓學生留在
該機構的干預措施，而非留在任何高等教育機構的干預措施）是一個重要區別。
- 在學 **(Enrolled)**：指在課程正常修業年限結束時仍註冊在學的學生 [6]。這些學生未在預期時間內畢業，但也未正式輟學；他們可能被視為學業延遲。此「在學」類別具有細微差別。這些學生尚未輟學，但也未按時畢業。此群體至關重要，因為他們代表了可能面臨挑戰、導致延遲畢業或最終輟學的學生。他們與明確的「畢業生」和「輟學生」有所不同。「在正常修業年限結束時仍註冊在學」[6] 的定義意味著他們已超過預期完成時間。此類別可能是旨在防止最終輟學或長期學習（這同樣具有成本）的干預措施的重點。部分研究移除此類別 [22] 雖然簡化了問題至二元分類，但也失去了這個重要的中間狀態。[11] 指出，預測「學業延遲的學生」（即在學學生）具有挑戰性。
- 畢業 **(Graduate)**：指在課程正常修業年限內成功完成學位的學生 [6]。

6.4 資料不平衡 此資料集的一個顯著特徵是類別間的嚴重不平衡，通常偏向某一類別（典

型為「畢業」)[1]。例如，[53](儘管是衍生資料集)的分布顯示「畢業」為多數，「輟學」佔 33%，其餘為「在學」。[59](一項使用此資料集的研究)報告「畢業」佔 49.93%，「輟學」佔 32.12%，「在學」佔 17.948%。[22](一個使用此資料的 GitHub 儲存庫)指出「畢業學生人數高於輟學學生人數」。這種「嚴重不平衡」(如眾多資料來源所述)不僅僅是一個數據特徵，更是使用此資料集的研究人員必須解決的核心方法論挑戰。標準演算法可能偏向多數類別，導致對少數類別(通常是「輟學」或「在學」)的預測效果不佳。這需要採用諸如 SMOTE [15]、成本敏感學習或選擇對不平衡具有魯棒性的演算法(例如某些提升方法 [11])。基於此資料集的預測模型的成功與否，在很大程度上取決於其處理這種不平衡的程度。這是任何使用此資料集的人都需要重點考慮的問題。**6.5** 部分研究中的目標變數修改 一些使用此資料集的研究修改了目標變數，例如，僅關注「畢業」和「輟學」類別，移除「在學」類別，以創建一個二元分類問題 [22]。

7. 在預測建模與教育研究中的應用

7.1 相關機器學習任務 此資料集主要用於多類別分類任務 [1]。部分研究將其轉換為二元分類問題 [22]，如前所述。**7.2** 常見應用的機器學習模型 文獻中，多種機器學習模型被應用於此資料集，其中以集成學習方法，特別是隨機森林 (Random Forest) 和各種提升演算法 (Boosting algorithms) 如 XGBoost、梯度提升 (Gradient Boost)、AdaBoost、LightGBM 及 CatBoost，因其通常表現出較優越的性能而備受關注 [11]。多項獨立研究(例如 [29]、[18]、[19]、[54]、[19]、[19]、[54])對於這些模型家族的效能得出了相似的結論。這表明這些演算法非常適合此資料集的特性(例如，混合類別型/數值型特徵、處理非線性關係、對某些雜訊的魯棒性，以及內在或增強的處理不平衡的能力)。這為新的研究人員提供了一個強而有力的起點。
其他常被提及的模型包括：

- 支援向量機 (Support Vector Machines, SVM) [10]
- 邏輯斯迴歸 (Logistic Regression) [10]
- 決策樹 (Decision Trees) [10]
- K-近鄰演算法 (K-Nearest Neighbors, KNN) [11]
- 類神經網路/深度學習 (Neural Networks / Deep Learning) [11]
- 其他集成方法 (堆疊法 Stacking, 投票法 Voting) [11]
- 樸素貝氏分類器 (Naive Bayes) [11]

7.3 特徵工程與選擇 特徵工程通常涉及類別變數的編碼和特徵縮放 [29]。特徵選擇技術，如相關性分析和基於模型(如隨機森林)的特徵重要性評估，也被廣泛使用 [13]。部分研究移除了與目標變數呈負相關的特徵，或使用如史萊姆黴菌演算法 (Slime Mould Algorithm) 進行特徵選擇 [22]。**7.4** 模型評估指標 常用的評估指標包括準確率 (Accuracy)、精確率 (Precision)、召回率 (Recall)、F1 分數 (F1-score) 和 ROC-AUC [5]。由於錯誤預測學生將畢業(偽陽性)的代價較高，部分研究特別強調精確率 [24]。混淆矩陣 (Confusion matrices) 則用於詳細呈現模型的預測效能 [29]。**7.5** 處理不平衡資料 預測模型的成功與否，在很大程度上取決於如何有效地管理類別不平衡。明確處理此問題(例如，使用 SMOTE 或選擇對不平衡敏感的演算法)的研究，通常能報告對少數類別更佳或更可靠的結果 [15]。資料集本身存在「嚴重不平衡」[1]。若忽略此問題，模型很可能僅透過預測多數類別來達到高準確率，卻在關鍵的少數類別(輟學、在學)上表現不佳。SMOTE (Synthetic Minority Over-sampling Technique) 等技術被廣泛採用 [15]。某些演算法(例如特定的提升演算法、帶有類別權重的隨機森林)本身能更好地處理不平衡問題 [26]。將「在學」類別移除以簡化為二元問題，也會改變不平衡的動態 [22]。在不平衡情境下，評估指標的選擇(例如，特定類別的 F1 分數、精確率/召

回率，而非整體準確率）也變得更加重要 [34]。**7.6** 建議的資料分割 原始專案使用了 80% 的訓練集和 20% 的測試集進行分割 [3]。其他研究也探索了不同的分割比例，例如從 50:50 到 90:10 [5]。

表 **2**：已應用機器學習模型及主要文獻發現摘要

此表彙整了使用此資料集進行機器學習建模的相關研究，提供了一個關於其實際應用和已達成效的即時概覽。它使讀者能夠快速比較不同建模方法及其在此特定資料集上的有效性。透過突顯各項研究的主要發現（如重要特徵），可以為新的研究提供指引，指出哪些變數持續顯示出預測能力。此表展示了已應用的技術範圍，從傳統模型到更先進的集成方法和深度學習，反映了教育資料探勘實踐的演進。它將來自許多不同來源的資訊（例如 [52]、[29]、[22] 等）合成為結構化且可比較的格式。

研究 (引用/來源 ID)	應用模型	主要報告效能指標 (例如：準確率、F1、輟學精確率)	主要發現/重要特徵	不平衡處理策略 (若有說明)
Arora [24]	18 種模型，包括神經網路、邏輯斯迴歸、KNN、AdaBoost、Gradient Boost、XGBOOST、隨機森林、SVM、決策樹、堆疊分類器	Tuned XGBoost 和 Stacking Classifier 表現最佳；強調精確率	重要特徵：「第二學期通過課程數」、「第一學期通過課程數」、「第一學期註冊課程數」、「第二學期成績」、「課程」、「第一學期評估課程數」	未明確說明，但討論了精確率的重要性
Ranga4all1 (GitHub) [29]	邏輯斯迴歸、決策樹、隨機森林、XGBoost	隨機森林因 ROC-AUC 較優而被選中	人口統計資訊、學術史、社會經濟指標、註冊詳情	提及處理類別不平衡 (model-train.ipynb)
Hamzaezzine (GitHub) [22]	KNN、邏輯斯迴歸、決策樹、隨機森林、SVM、樸素貝氏	SVM 在準確率、精確率、召回率和 F1 分數方面表現最佳	移除與目標呈負相關的特徵	移除了「在學」類別
Ridwan et al. (JECA) [5]	XGBoost	80:20 分割下，精確率 88%，F1 分數 81%	使用 StandardScaler 進行正規化	多種訓練/測試分割比例
Alecngo (GitHub) [37]	邏輯斯迴歸、KNN、決策樹、隨機森林、SVM、深度學習	深度學習將準確率從 0.72 提升至 0.92	特徵工程（消除離群值、刪除重複特徵）	Dropout 和 EarlyStopping (深度學習)

Martins et al. (2023) [26]	五種機器學習演算法 (包括隨機森林)	隨機森林表現佳；第一學期末預測效果好	社會人口統計學、宏觀經濟、學術資訊	處理資料不平衡的策略 (資料層面或演算法層面)
Martins et al. (2021) [18]	邏輯斯迴歸、SVM、決策樹、隨機森林、提升演算法	提升演算法反應較好，但少數類別識別仍有不足	入學時的學術路徑、人口統計學、社會經濟因素	測試了促進類別平衡的合成過採樣方法
JCT Journal [13]	未指定具體模型，但提及分類器	未提供具體數值	重要特徵：「第二學期通過課程數」、「第一學期通過課程數」、「第二學期成績」、「第一學期成績」、「學費是否繳清」	
Fahd et al. (JHETP) [15]	邏輯斯迴歸、隨機森林	隨機森林優於邏輯斯迴歸	學術表現是成功的顯著預測變數	SMOTE

儘管高性能模型如複雜集成或深度學習 [11] 在預測方面很有價值，但更簡單的模型如決策樹 [57] 或帶有特徵重要性分析的邏輯斯迴歸 [34] 可以提供更好的可詮釋性。這對於理解學生為何可能面臨風險以及為制定可行的干預措施至關重要。最終目標通常不僅僅是預測，而是理解和干預。高度複雜的「黑箱」模型可能提供良好的預測，但對於驅動因素的洞察卻很少。[57] 提到決策樹提高了詮釋性。[52] 和 [29] 討論了特徵重要性。這表明需要在預測能力與提取教育者和管理者可以使用的可行見解之間取得平衡。一些研究 [31] 甚至正在探索新穎的特徵選擇方法以提高可詮釋性。

8. 現有研究的關鍵因素與洞見

8.1 持續重要的預測因子

綜合多項研究，以下因素被認為對預測學生輟學與學業成功具有顯著影響：

- 前期學業表現：先前學期的學業表現，特別是「第二學期通過課程數 (Curricular units 2nd sem (approved))」和「第一學期通過課程數 (Curricular units 1st sem (approved))」，常被認為是最具影響力的特徵 [9]。這表明早期的學術成功或失敗會產生強烈的「動量」，顯著影響學生堅持學業並最終畢業的可能性。早期表現優異可能會建立信心、參與度，並為未來課程奠定基礎。相反，早期困境（未通過課程、成績不佳）可能導致沮喪、落後，並增加輟學的可能性。這突顯了第一學年學習的關鍵性。因此，專注於支持學生度過最初幾個學期的干預措施可能極具影響力。
- 學期成績：第一和第二學期課程單元的成績（「第一學期成績 (Curricular units 1st sem (grade))」、「第二學期成績 (Curricular units 2nd sem (grade))」）也同樣非常重要 [13]。
- 財務狀況：「學費是否繳清 (Tuition fees up to date)」是一個強有力的預測因子，尤其對輟學而言 [9]。這表明財務穩定性是學術成功的關鍵促成因素。面臨財務困難（負債、未繳學費）的學生輟學風險顯著較高。財務壓力可能會分散學習注意力，迫使學生過度

工作，或因無力支付費用而退學。獎學金則減輕了這種壓力。這意味著財務支持系統和靈活的支付方案對於學生續讀而言，可能與學術支持同等重要。
- 獎學金狀況：「是否持有獎學金 (Scholarship holder)」也扮演一定角色 [9]。
- 其他學術因素：如「第一學期註冊課程數 (Curricular units 1st sem (enrolled))」和「第一學期評估課程數 (Curricular units 1st sem (evaluations))」也顯示出其重要性 [52]。
- 人口統計學因素：「入學年齡 (Age at enrollment)」可能是一個因素，年齡較大的學生有時風險較高 [16]。「性別 (Gender)」和「婚姻狀況 (Marital status)」在不同研究中顯示出不同程度的重要性，有時表明男性或已婚學生風險較高 [9]。
- 先前學歷：「先前學歷 (Previous qualification)」有時相較於當前學業表現，其預測能力較低 [9]。雖然直觀上人們可能期望先前的學術成就具有高度預測性，但一些關於此資料集的研究發現「先前學歷成績 (Previous qualification (grade))」或「先前學歷 (Previous qualification)」的重要性低於當前學期的表現 [9]。這可能意味著：(a) 向高等教育的過渡是一個重大轉變，學生如何適應並在
新環境中表現，迅速變得比他們過去的中學成績更為關鍵。(b) 先前學歷的多樣性和潛在的不可比性（如第 4 節洞見 3 所述）在簡單建模時可能會削弱其預測能力。(c) 高等教育機構內部的即時壓力和表現（學術負荷、教學風格、社會融合）迅速掩蓋了先前的成就。這並非意味著先前的成就無關緊要，而是其直接的預測信號可能弱於更近期的表現指標。

8.2 主要引用學術論文的洞見

- **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."** [3]。
  - 這篇基礎性論文使用此資料集建立了分類模型，測試了邏輯斯迴歸、SVM、決策樹和隨機森林等演算法 [10]。
  - 研究發現提升演算法對此任務反應較好，但在處理少數類別時仍有困難 [18]。
  - 強調使用入學資料（學術路徑、人口統計學、社會經濟因素）進行早期識別 [18]。
  - 指出了嚴重的類別不平衡問題 [18]。
  - 未來的研究方向建議加入第一年的表現數據（此資料集版本已包含）[18]。
- **Realinho, V., Machado, J., Baptista, L., Martins, M.V. (2022). "Predicting Student Dropout and Academic Success." *Data*, *7*, 146.** [6]。
  - 這篇 MDPI 論文詳細描述了資料集本身、其創建過程、屬性及預處理方法 [6]。
  - 重申了使用機器學習進行預測，作為學習分析工具一部分的目標 [6]。
  - 提及在其更廣泛的專案中使用了 RF、XGBoost、LGBM 和 CatBoost [10]。
- **Martins, M.V., Baptista, L., Machado, J., Realinho, V. (2023). "Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education." *Applied Sciences*.** [26]。
  - 評估了在第一學年期間進行預測的最佳時機，使用了資料集的不同「版本」（入學時數據、第一學期中數據、第一學期末數據）。

- 研究發現，當部分學業表現數據可用時，在第一學期末能獲得最佳的預測結果。
- 隨機森林表現良好，尤其是在採用處理不平衡數據的策略時。
- 旨在建立一個可推廣至該理工大學內各個學位的系統。

9. 資料集使用建議

9.1 承認並處理類別不平衡 由於固有的類別不平衡，研究人員必須採用適當的策略，例如 SMOTE、成本敏感學習、專為不平衡設計的集成方法，以及合適的評估指標（如 F1 分數、少數類別的精確率/召回率）[1]。
9.2 仔細的特徵工程與選擇

- 適當編碼類別變數（許多變數雖以數值編碼，但本質上是名目型或序數型）。需考慮 Course 或 Previous Qualification 等變數的高基數問題 [3]。
- 探索特徵之間的交互作用，特別是學術、社會經濟和人口統計變數之間的交互作用。
- 利用模型的特徵重要性來理解驅動因素，但同時也應結合領域知識。

9.3 交叉驗證策略 採用穩健的交叉驗證方法，考慮到類別不平衡，可能需要使用分層交叉驗證，以獲得可靠的模型性能評估 [5]。**9.4** 可詮釋性 當目標是制定干預措施時，除了預測準確性外，應優先考慮提供可詮釋性的模型（例如決策樹、帶有係數分析的邏輯斯迴歸、用於複雜模型的 SHAP 值 ）[34]。**9.5** 結果的脈絡化 需注意此資料來自葡萄牙單一理工學院。將研究結果推廣至其他情境時應謹慎考慮並加以討論 [6]。

9.6 倫理考量
注意數據和模型中潛在的偏見，尤其是在人口統計學和社會經濟特徵方面。確保預測模型用於支持學生，而非不公平地標籤化或懲罰他們。
9.7 分階段預測 考慮 Martins 等人 (2023) [26] 關於預測最佳時機的研究結果（第一學期末對於早期預測是有效的）。此資料集允許在不同階段（僅入學時、第一學期後、第二學期後）進行建模。
儘管此資料集提供了截至第二學期的數據，但學生的學習旅程仍在繼續。為了更全面地了解情況，未來的數據收集可以進一步擴展。然而，對於早期預測而言，此資料集非常適用。目前的資料集是早期預測的一個快照；真正的縱向追蹤需要隨時間推移收集更多的數據點。此資料集並非為此目的而設計的缺陷，而是更廣泛的學生成功研究所需考慮的因素。

預測模型只有在能夠引導有效干預時才有用。相關建議應強調將模型發現（例如關鍵風險因素）轉化為機構可以實施的實際策略。這與該資料集旨在「制定」並「實施支持學生的策略」[3] 的目的相符。這意味著使用此數據的研究成果，理想情況下應不僅僅是一篇研究論文，更應為實踐提供資訊。這也與學習分析工具的開發緊密相連 [6]。納入宏觀經濟變數（失業率、通貨膨脹率、GDP）[6] 是一個獨特的特點。研究人員應探討這些因素是作為影響所有學生的普遍壓力源，還是與個別學生特徵（例如，來自較低社會經濟背景的學生更容易受到高失業率的影響）產生交互作用。宏觀經濟狀況 [6] 可能會造成經濟壓力或機會的背景，從而影響輟學決策（例如，工作機會成為「拉力因素」[24]）。這些因素較少涉及個別學生屬性，而更多地關乎環境。分析其影響，可能將其作為模型中的交互項，可以為了解更廣泛的經濟氣候如何影響學生續讀率帶來有趣的見解。
10. 結論與未來方向

10.1 價值總結「預測學生輟學與學業成功」資料集作為一個經過預處理的綜合資源，對於研究學生輟學與學業成功，尤其是在歐洲高等教育背景下的早期預測，具有重要價值。它使得研究人員能夠深入探討人口統計學、社會經濟及學術因素之間複雜的交互作用。此資料集相對較新（2021 年捐贈），但已激發了大量研究（眾多引用和 Kaggle 筆記本證明了這一點）[5]。這突顯了公開此類數據的影響力，並強化了開放數據在科學研究中的價值。

10.2 基於此資料集的未來研究潛力

- 進一步探索先進的特徵工程技術。
- 對更新的機器學習演算法或深度學習架構進行比較研究。
- 更深入地分析「在學」(Enrolled) 類別（即學業延遲的學生）。
- 調查源自此數據的預測模型中的公平性與偏見問題。
- 探索遷移學習：基於此資料集訓練的模型，經過一定的重新校準後，是否能適應其他（可能規模較小的）機構資料集？

儘管許多研究應用了不同的模型，但仍需要系統性地複製研究結果，並測試模型在不同數據分割、預處理變體（若有）和超參數設定下的魯棒性，以便對此類數據的最佳方法建立更明確的結論。許多研究報告了各自的最佳模型（例如，[52] 發現堆疊分類器，[29] 發現隨機森林，[22] 發現 SVM）。雖然這些都很有價值，但在相同條件下進行更系統的比較，或專注於複製先前結果的研究，將有助於該領域更深入的理解。建議的 80/20 分割 [3] 提供了一個基準，但仍存在差異。

10.3 對教育資料探勘的更廣泛啟示

- 持續需要高質量、文檔完善的教育資料集。嚴謹的預處理和清晰的文檔記錄（儘管前面提到屬性計數存在微小不一致）[3] 對於可以產出的研究品質有顯著貢獻。在清理上花費更少資源意味著有更多資源用於進階分析。這是一個直接的因果關係：更好的輸入數據有助於更好、更有效率的研究產出。此資料集是資料集創建者如何支持研究社群的一個良好範例。
- 將預測建模與可操作的、合乎倫理的干預措施相結合的重要性。
- 在模型性能與可詮釋性之間取得平衡的持續挑戰。

11. 學術引用文獻

以下列出與此資料集相關的主要學術論文，資訊來源為 UCI 頁面及 MDPI 論文：

- 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. DOI: 10.1007/978-3-030-72657-7_16 [3]
- Realinho, V.; Machado, J.; Baptista, L.; Martins, M.V. (2022). "Predicting Student Dropout and Academic Success". *Data*, *7*, 146. DOI: 10.3390/data7110146 [6]
- UCI 資料集引用：Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and

Baptista,Luís. (2021). Predict Students' Dropout and Academic Success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89 [38]

引用的著作

1. Predict Students' Dropout and Academic Success - UCI Machine ..., 檢索日期：5月 6, 2025, https://archive.ics.uci.edu/dataset/697/predict%2Bstudents%2Bdropout%2Band%2Bacademic%2Bsuccess
2. Datasets - UCI Machine Learning Repository, 檢索日期：5月 6, 2025, https://archive.ics.uci.edu/datasets?search=&Keywords=Imbalanced%20classes
3. Predict Students' Dropout and Academic Success - UCI Machine Learning Repository, 檢索日期：5月 6, 2025, https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success
4. Students' Dropout and Academic Success - Kaggle, 檢索日期：5月 6, 2025, https://www.kaggle.com/datasets/satyajeetbedi/students-dropout-and-academic-success
5. Predict Students' Dropout and Academic Success with XGBoost - Journal of Education and Computer Applications, 檢索日期：5月 6, 2025, https://jeca.aks.or.id/jeca/article/download/13/6/61
6. Predicting Student Dropout and Academic Success - MDPI, 檢索日期：5月 6, 2025，https://www.mdpi.com/2306-5729/7/11/146
7. Predicting Student Dropout and Academic Success - Semantic Scholar, 檢索日期：5月 6, 2025, https://pdfs.semanticscholar.org/6f6b/c57232d584b79fca7852c41f77b103ca9929.pdf
8. Predicting Student Dropout and Academic Success - OUCI, 檢索日期：5月 6, 2025，https://ouci.dntb.gov.ua/en/works/9GrDmkZl/
9. EXPLORING THE DETERMINANTS OF STUDENTS' DROPOUT IN HIGHER EDUCATION A Machine Learning Approach - ULB : Dok - Universität Innsbruck, 檢索日期：5月 6, 2025，https://ulb-dok.uibk.ac.at/ulbtirolhs/download/pdf/10391315
10. CLASSIFICATION OF STUDENTS' DROPOUT STATUS USING THE RANDOM FOREST METHOD - International Journal Of Eurasia Social Sciences, 檢索日期：5月 6, 2025，https://www.ijoess.com/DergiPdfDetay.aspx?ID=4507
11. Predicting student dropout using multiclass classification A comparative study on ensemble learning methods - Tilburg University, 檢索日期：5月 6, 2025, http://arno.uvt.nl/show.cgi?fid=181524
12. Student retention analysis - Theseus, 檢索日期：5月 6, 2025, https://www.theseus.fi/bitstream/10024/857738/2/Gautam_Lok.pdf
13. UNDER GRADUATE STUDENT DROPOUT PREDICTION USING MACHINE LEARNING - A JOURNAL OF COMPOSITION THEORY(JCT), 檢索日期：5月 6, 2025，https://jctjournal.com/wp-content/uploads/23-july2023.pdf
14. INTELLIGENT METHODS IN ENGINEERING SCIENCES Predicting Student Dropout Using Machine Learning Algorithms, 檢索日期：5月 6, 2025,

https://imiens.org/index.php/imiens/article/download/62/38/469
15. A Comparative Study of Machine Learning Techniques for College Student Success Prediction - Article Gateway, 検索日期：5月 6, 2025, https://articlegateway.com/index.php/JHETP/article/download/6764/6391
16. Dropout in Higher Education: Data Visualization - RPubs, 検索日期：5月 6, 2025, https://rpubs.com/ncdea/dropout-dv
17. shivamsingh96/Predict-students-dropout-and-academic-success - GitHub, 検索日期：5月 6, 2025, https://github.com/shivamsingh96/Predict-students-dropout-and-academic-success
18. Early Prediction of student's Performance in Higher Education: A Case Study | CoLab, 検索日期：5月 6, 2025, https://colab.ws/articles/10.1007%2F978-3-030-72657-7_16
19. Early Prediction of Student's Performance in Higher Education: A Case Study | Request PDF, 検索日期：5月 6, 2025, https://www.researchgate.net/publication/351066139_Early_Prediction_of_Student's_Performance_in_Higher_Education_A_Case_Study
20. (PDF) DropWrap: A Neural Network Based Automated Model for Managing Student Dropout, 検索日期：5月 6, 2025, https://www.researchgate.net/publication/390687660_DropWrap_A_Neural_Network_Based_Automated_Model_for_Managing_Student_Dropout
21. Predict students' dropout and academic success - Kaggle, 検索日期：5月 6, 2025, https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention
22. hamzaezzine/Predict-students-dropout-and-academic-success-using-machine-learning-algorithms - GitHub, 検索日期：5月 6, 2025, https://github.com/hamzaezzine/Predict-students-dropout-and-academic-success-using-machine-learning-algorithms
23. Predict students dropout, academic success - Kaggle, 検索日期：5月 6, 2025, https://www.kaggle.com/datasets/naveenkumar20bps1137/predict-students-dropout-and-academic-success
24. Predicting Students Academic Success and Dropout Using Supervised Machine Learning - international journal of scientific study, 検索日期：5月 6, 2025, https://www.ijss-sn.com/uploads/2/0/1/5/20153321/14_ijss_sep_23_oa12_-_2023.pdf
25. Datasets - UCI Machine Learning Repository, 検索日期：5月 6, 2025, https://archive.ics.uci.edu/datasets?search=&Keywords=Academic%20performance
26. Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education, 検索日期：5月 6, 2025, https://www.researchgate.net/publication/369937622_Multi-Class_Phased_Prediction_of_Academic_Performance_and_Dropout_in_Higher_Education
27. Students' Dropout and Academic Success Prediction - Kaggle, 検索日期：5月 6, 2025, https://www.kaggle.com/datasets/susanketsarkar/students-dropout-and-academ

ic-success-prediction
28. EDA and Prediction of Student Academic Success - Kaggle, 検索日期：5月 6, 2025 , https://www.kaggle.com/code/paulandrewpaglinawan/eda-and-prediction-of-student-academic-success
29. Predict students' dropout and academic success - GitHub, 検索日期：5月 6, 2025 , https://github.com/ranga4all1/student-dropout-and-success-prediction
30. Predict Student's Dropout and Academic Success ML - Kaggle, 検索日期：5月 6, 2025, https://www.kaggle.com/code/naveenkumar20bps1137/predict-student-s-dropout-and-academic-success-ml
31. A novel approach to mitigate academic underachievement in higher education: Feature selection, classifier performance, and interpretability in predicting student performance - Science Gate, 検索日期：5月 6, 2025, https://www.science-gate.com/IJAAS/2024/V11I5/1021833ijaas202405015.html
32. Student Dropout Analysis for School Education - Kaggle, 検索日期：5月 6, 2025, https://www.kaggle.com/code/jeevabharathis/student-dropout-analysis-for-school-education
33. University-Success-Prediction-Model - Kaggle, 検索日期：5月 6, 2025, https://www.kaggle.com/code/gabrielebosi/university-dropout-prediction-model
34. Exploring Individual Feature Importance in Student Persistence Prediction - ResearchGate, 検索日期：5月 6, 2025, https://www.researchgate.net/publication/373863254_Exploring_Individual_Feature_Importance_in_Student_Persistence_Prediction
35. Students Dropout and Academic Success Dataset - Kaggle, 検索日期：5月 6, 2025 , https://www.kaggle.com/datasets/mahwiz/students-dropout-and-academic-success-dataset
36. Datasets - UCI Machine Learning Repository, 検索日期：5月 6, 2025, https://archive.ics.uci.edu/datasets?search=&Keywords=Classification
37. alecngo/Predict-students-dropout-and-academic-success - GitHub, 検索日期：5月 6, 2025, https://github.com/ahnngo/Predict-students-dropout-and-academic-success
38. noahk587/Student-Dropout-Prediction: A personal project where logistic regression is used to predict if a student dropped out. - GitHub, 検索日期：5月 6, 2025, https://github.com/noahk587/Student-Dropout-Prediction
39. Predict Students' Dropout and Academic Success with XGBoost, 検索日期：5月 6, 2025, https://jeca.aks.or.id/jeca/article/view/13
40. Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education, 検索日期：5月 6, 2025, https://www.mdpi.com/2076-3417/13/8/4702
41. Применение учебной аналитики в высшем образовании: датасеты, методы и инструменты | Дюличева, 検索日期：5月 6, 2025, https://vovr.elpub.ru/jour/article/view/4980
42. Application of Learning Analytics in Higher Education: Datasets, 検索日期：5月 6, 2025, https://ouci.dntb.gov.ua/en/works/73Nn8r89/

43. A Data Feature Extraction Method Based on the NOTEARS Causal Inference Algorithm, 検索日期：5月 6, 2025, https://www.researchgate.net/publication/372617438_A_Data_Feature_Extraction_Method_Based_on_the_NOTEARS_Causal_Inference_Algorithm
44. Análisis comparativo de Técnicas de Machine Learning para la predicción de casos de deserción universitaria - SciELO Portugal, 検索日期：5月 6, 2025, https://scielo.pt/scielo.php?script=sci_arttext&pid=S1646-98952023000300084&lng=pt&nrm=iso&tlng=es
45. Применение учебной аналитики в высшем образовании: датасеты, методы и инструменты Текст научной статьи по специальности - КиберЛенинка, 検索日期：5月 6, 2025, https://cyberleninka.ru/article/n/primenenie-uchebnoy-analitiki-v-vysshem-obrazovanii-datasety-metody-i-instrumenty
46. Prof. Dr. Valentim Alberto Correia Realinho | Author - SciProfiles, 検索日期：5月 6, 2025, https://sciprofiles.com/profile/vrealinho
47. Early Prediction of student's Performance in Higher Education: A Case Study - OUCI, 検索日期：5月 6, 2025, https://ouci.dntb.gov.ua/en/works/4wJjdLJ7/
48. Forecasting Student Dropout and Academic Achievement (UCI Machine Learning Repository) | Spreadsheet Download | Gigasheet, 検索日期：5月 6, 2025, https://www.gigasheet.com/sample-data/predict-students-dropout-and-academic-success
49. A Comparative Study of Machine Learning Techniques for College Student Success Prediction - Article Gateway, 検索日期：5月 6, 2025, https://articlegateway.com/index.php/JHETP/article/download/6764/6391/11678
50. repositorio.ucp.pt, 検索日期：5月 6, 2025, https://repositorio.ucp.pt/bitstream/10400.14/44855/1/203589203.pdf
51. Predict students' dropout and academic success - Kaggle, 検索日期：5月 6, 2025, https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/discussion
52. Predicting Students Academic Success and Dropout Using Supervised Machine Learning, 検索日期：5月 6, 2025, https://www.researchgate.net/publication/384055745_Predicting_Students_Academic_Success_and_Dropout_Using_Supervised_Machine_Learning
53. Classification with an Academic Success Dataset - Suraj Wate, 検索日期：5月 6, 2025, https://surajwate.com/blog/classification-with-an-academic-success-dataset/
54. Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study - ResearchGate, 検索日期：5月 6, 2025, https://www.researchgate.net/publication/377170227_Supervised_machine_learning_algorithms_for_predicting_student_dropout_and_academic_success_a_comparative_study
55. Student Dropout Prediction for University with High Precision and Recall - ResearchGate, 検索日期：5月 6, 2025, https://www.researchgate.net/publication/370965520_Student_Dropout_Prediction_for_University_with_High_Precision_and_Recall

56. End-to-End MLOps Pipeline: A Comprehensive Project | GeeksforGeeks, 検索日期：5月 6, 2025, https://www.geeksforgeeks.org/end-to-end-mlops-pipeline-a-comprehensive-project/
57. Predicting Academic Success of College Students Using Machine Learning Techniques, 検索日期：5月 6, 2025, https://www.mdpi.com/2306-5729/9/4/60
58. AlphaML: A clear, legible, explainable, transparent, and elucidative binary classification platform for tabular data - PMC, 検索日期：5月 6, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10801203/
59. Dropout and Graduation in Higher Education: CHAID Analysis - DergiPark, 検索日期：5月 6, 2025, https://dergipark.org.tr/en/pub/eku/issue/85904/1287393
60. Full Schedule - SCCUR 2023, 検索日期：5月 6, 2025, https://sccur2023.sched.com/list/descriptions/
61. Student Performance - UCI Machine Learning Repository, 検索日期：5月 6, 2025, https://archive.ics.uci.edu/dataset/320/student+performance