

أمامكم بيانات أكاديمية لجامعة افتراضية، تحتوي على 395 صفًا و33 عمودًا. البيانات تشمل تفاصيل الطلاب مثل المدرسة، الجنس، العمر، نوع السكن، حجم العائلة، المستوى التعليمي للأبوين، وظائفهم، أسباب اختيار المدرسة، الوصي، زمن السفر، وقت الدراسة، عدد الإخفاقات، الدعم المدرسي، الدعم العائلي، الدروس المدفوعة، النشاطات، الحضانة، الرغبة في التعليم العالي، توفر الإنترنت، الحالة العاطفية، العلاقات الأسرية، وقت الفراغ، الخروج، استهلاك الكحول، الحالة الصحية، الغياب، والدرجات g1, g2, g3.

الأعمدة تحتوي على بيانات كمية (مثل: العمر، زمن الدراسة، الغياب، الدرجات) وبيانات نوعية (مثل: الجنس، المدرسة، الدعم الأسري، الحالة العاطفية). بعض الأعمدة الخاصة بالدرجات مخزنة ك object وتحتاج تحويل قبل التحليل.

البيانات المقدمة:

<https://archive.ics.uci.edu/dataset/320/student+performance>

الأهداف:

1. تنظيف البيانات وتجهيزها للتحليل.
2. استكشاف البيانات (Exploratory Data Analysis – EDA) وفهم الأنماط الأساسية.
3. تطبيق أساليب Data Mining على البيانات، بما في ذلك التصنيف، الانحدار، التجميع، قواعد الترابط، وتحليل البيانات الزمنية.

المطلوب:

الجزء 0: Know Your Data

قبل البدء بأي تحليل أو نموذج، تعرف على البيانات:

1. عرض الأعمدة وأنواعها.
2. التحقق من القيم المفقودة.
3. تحويل الأعمدة التي تخزن الأرقام ك object إلى أرقام.

4. الكشف عن القيم الشاذة (مثل درجات أو حضور غير منطقي).

5. الفهم العام للبيانات:

- عدد الصفوف والأعمدة.
- توزيع الطلاب حسب المدرسة والجنس.
- تحليل المتوسط العام للدرجات G1, G2, G3.
- دراسة العلاقة بين الدعم الأسري والمدرسي ومستوى الأداء.
- مقارنة أداء الطلاب حسب الجنس أو نوع السكن (Urban/Rural).

الجزء 1: Data Cleaning

- إزالة أو معالجة القيم المفقودة.
- تحويل الأعمدة الخاصة بالدرجات (G1, G2, G3) إلى أرقام.
- معالجة القيم الشاذة (مثل درجات سالبة أو أكبر من 20).
- التأكد من منطقية القيم مثل العمر، عدد الإخفاقات، الغياب.

الجزء 2: Exploratory Data Analysis

- عرض ملخص إحصائي للأعمدة الرقمية (mean, median, std...).
- رسم توزيعات الدرجات (G1, G2, G3) والغياب والعمر.
- تحليل فروقات الأداء بين الجنسين والمدارس ونوع السكن.
- تحليل العلاقة بين الحضور (absences) والدرجات النهائية (G3).
- دراسة تأثير الدعم المدرسي (schoolsup) والعائلي (famsup) على الأداء.

الجزء3: Data Mining Tasks

-1 Classification

التنبؤ بـ مستوى الأداء (Excellent / Good / Poor)

بناءً على خصائص مثل:

studytime, failures, absences, schoolsup, famsup, internet, health, G1, G2.

-2 Regression

التنبؤ بـ الدرجة النهائية (G3)

من الخصائص التالية:

G1, G2, studytime, absences, failures, Dalc, Walc, health.

-3 Clustering

تجميع الطلاب حسب السلوك الدراسي والخصائص المشتركة مثل:

studytime, absences, failures, goout, freetime, G3.

وتحليل خصائص كل مجموعة.

-4 Time Series Analysis

في حال اعتبار G1, G2, G3 مراحل زمنية (مثل فصل أول، ثاني، نهائي):

- تحليل تطور أداء الطالب عبر الفصول.
- التنبؤ بالدرجة النهائية بناءً على الأداء السابق.

الجزء4: Report

- وثق جميع خطواتك: تنظيف البيانات، التحليل الاستكشافي، الرسوم البيانية.
- عرض نتائج التصنيف، الانحدار، التجميع، قواعد الترابط، والتحليل الزمني.

- استنتاجات:

مراحل زمنية (مثل فصل أول، ثاني، نهائي): G1, G2, G3 في حال اعتبار

تحليل تطور أداء الطالب عبر الفصول.

التنبؤ بالدرجة النهائية بناءً على الأداء السابق.

يمكنك استخدام مكتبات Python مثل:

pandas, numpy, matplotlib, seaborn, scikit-learn, mlxtend.

وأيضا تسليم الشغل كـ Jupyter Notebook

ورفعه على github.

Nora Hussain