

# Laporan Proyek UTS - IBDA3111

Calvin Institute of Technology

Semester ganjil 2022/2023



Oleh

Victor Chendra / 202000338 / IT & Big Data Analytics

## Students Performance in Exams Dataset

### A. Latar belakang

Dalam mata kuliah ini, yaitu *IBDA3111 Prapemrosesan & Rekayasa data*, kami diberikan sebuah proyek UTS dengan mencari dataset sendiri untuk diolah lalu dianalisa atau melakukan pemodelan.

Adapun dataset yang saya pilih untuk proyek ini adalah dataset yang berjudulkan Students Performance in Exams Dataset yang diambil dari situs [kaggle.com](https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams).

Berikut adalah tautan dataset:

<https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams>

*Students Performance in Exams* dataset adalah sebuah dataset yang menggambarkan tentang peforma ujian siswa. Dalam dataset berisikan tentang nilai-nilai ujian siswa di sekolah. Adapun tabel-tabel yang terdapat di dalam dataset.

gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
string	string	string	string	string	integers	integers	integers
male	group A	high school	standard	completed	67	67	63
female	group D	some college	standard	none	59	60	50
female	group C	high school	standard	none	63	73	66
female	group C	bachelor's degree	standard	none	51	64	63
male	group D	high school	free/reduced	completed	82	83	77

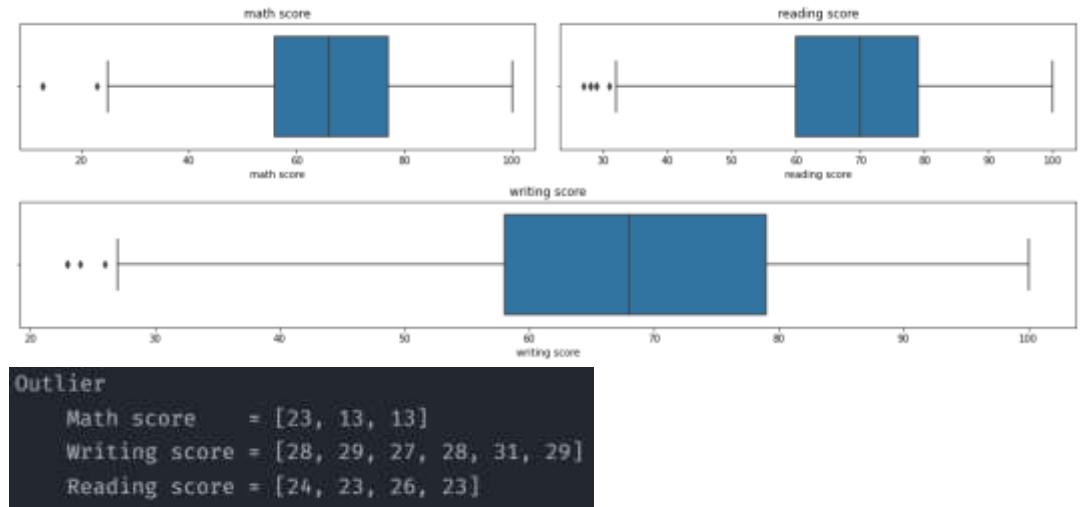
Terdapat 8 kolom, 5 string 3 integers dan 1000 baris data. Dataset ini dapat digunakan untuk analisa statistika maupun membangun pemodelan untuk menebak nilai ujian seorang siswa. Dataset ini juga memberikan gambaran, apakah pengaruh latar belakang pendidikan orang tua, *test preparation course*, dan lain-lain terhadap kinerja ujian siswa.

Adapun masalah yang ingin saya ingin selesaikan secara *general* adalah:

- i. How effective is the test preparation course?
- ii. Which major factors contribute to test outcomes?
- iii. What would be the best way to improve student scores on each test?
- iv. What patterns and interactions in the data can you find?

## B. Data

- Pemrosesan data yang saya lakukan dalam proyek ini yaitu, 2 metode *data cleaning* dan 2 metode *feature selection* yang mana merupakan kriteria yang telah ditentukan dalam proyek ini. Dua metode *data cleaning* tersebut adalah:
  - Delete duplicated rows:*  
Terdapat 1 baris data yang terduplicasi
  - Identify outlier and imputation the outlier:*  
Terdapat 13 data pencilan dalam dataset



Data pencilan tersebut, diimputasi ke batas bawah mereka masing-masing.

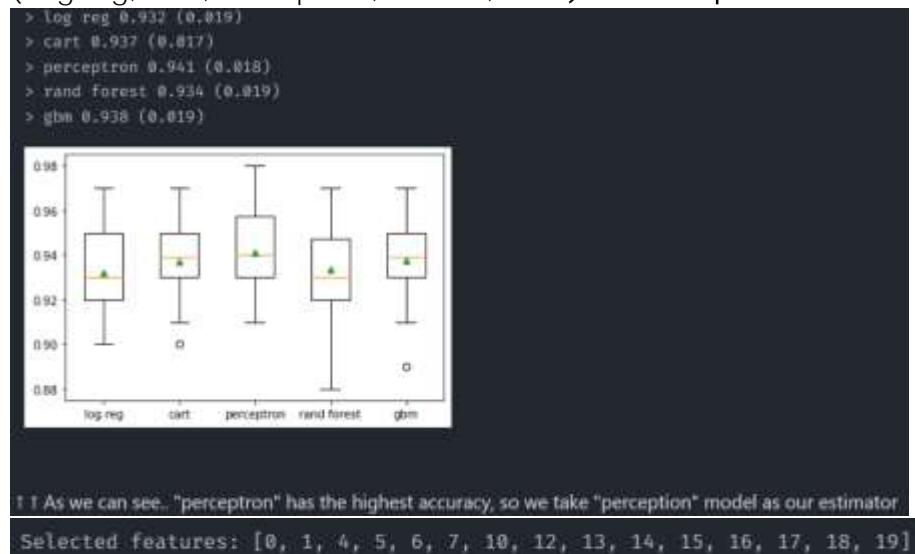
Kemudian untuk 2 metode feature selection:

- RFE
  - Chi-square
- Pendekatan yang saya lakukan dalam proyek ini adalah analisis statistika dan visualisasi. Alasan saya memilih analisis statistika untuk proyek/dataset ini adalah karena menurut saya analisa statistika lebih menyenangkan dan lebih menunjukkan sebuah hasil/gambaran dataset tersebut. Walaupun sebenarnya bisa melakukan pemodelan yaitu menebak/prediksi atau klasifikasi (*pass or fail*), tetapi kurang tepat saja dilakukan pada dataset ini dibandingkan dengan analisis statistika.

## C. Solusi

- Langkah-langkah yang saya lakukan dalam mengerjakan proyek ini:
  - Import library yang dibutuhkan
  - Load and check dataset

- iii. Data cleaning → 1. Delete duplicated rows → 2. Identify outlier & impute to its lower bound value
- iv. Feature selection (dataset after data cleaning)
  - 1. (RFE) Add new column “average” → one hot encoding the input variables → change the output variables to PassStatus (if the value is  $\geq 60$ , it's mean pass or 1 else 0 fail) → then we select the best RFE estimator from 5 RFE models estimator (LogReg, Cart, Perceptron, RanFor, GBM) it's Perceptron.



1.1 As we can see, "perceptron" has the highest accuracy, so we take "perception" model as our estimator

Selected features: [0, 1, 4, 5, 6, 7, 10, 12, 13, 14, 15, 16, 17, 18, 19]

Selected column according RFE model Perceptron

gender_Female	gender_Male	race/ethnicity_group_C	race/ethnicity_group_D	race/ethnicity_group_E	parental_level_of_education_associates_degree	parental_level_of_education_master_s_degree	parental_level_of_education_some_high_school
0	0	1	0	0	0	0	0
1	1	0	0	1	0	0	1
lunch_free/reduced	lunch_standard	test_preparation_course_completed	test_preparation_course_none	math_PassStatus	read_PassStatus	write_PassStatus	average
0	1	1	0	1	1	1	1
1	0	0	1	0	0	0	0

## 2. Chi-square

Chi-square output is the score of each input variables. → Sort it but keep the original index → then display the final dataset after features selection chi-square (best of 15 out of 20 features)

Selected column according chi-square score (best of 15 out of 20 features)

parental level of education, some college	race/ethnicity group E	parental level of education, master's degree	parental level of education, bachelor's degree	parental level of education, associate's degree	race/ethnicity group D	race/ethnicity group C	test preparation course, none
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1
2	1	1	0	0	0	0	1
parental level of education, some high school	test preparation course completed	lunch, standard	lunch, free/reduced	read_PassStatus	math_PassStatus	write_PassStatus	average
0	1	1	0	1	1	1	1
1	0	0	1	0	0	0	0
0	0	0	1	1	0	0	0

- v. Answering the problem define in description

Untuk menjawab masalahnya, pertama kita melakukan visualisasi dataset untuk diamati lalu mengambil kesimpulan.

#### D. Kesimpulan

- Dataset diperoleh dari [kaggle.com](https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams)  
<https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams>
- Cara memperoleh dataset: mengunduh dari website
- Terdapat 2 metode *cleaning data*: Delete duplicated rows, Identify outlier & Imputation.
- Terdapat 2 metode feature selection: RFE, Chi-square
- Pendekatan analisis statistika
- Hasil observasi dataset:
  - Students with a higher parental's degrees have more chances to achieve a higher scores
  - Based on gender, female have higher score average than male
  - The females dominate the higher scores in reading and writing
  - Male outperform female in ONLY math
  - Along this dataset female and male students are equally distributed
  - "Lunch" is quite influential in performing student's exams score
- Special thanks untuk beberapa sumber inspirasi:
  - <https://medium.com/@srajaninnov/students-performance-in-exams-data-analysis-19ca93fccd37>
  - <https://www.kaggle.com/code/ibrahimalmanaseer/student-performance>