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.
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:

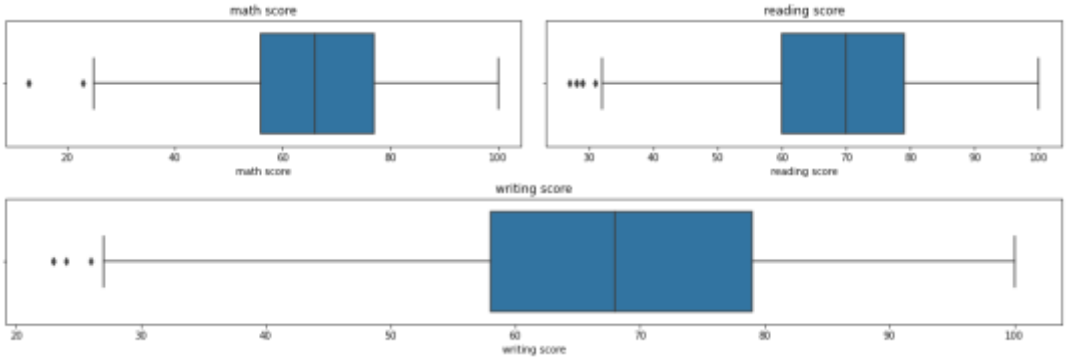
i. How effective is the test preparation course?
ii. Which major factors contribute to test outcomes?
iii. What would be the best way to improve student scores on each test?
iv. What patterns and interactions in the data can you find?

B. Data

a. 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* tersbut adalah:

i. *Delete duplicated rows*:
Terdapat 1 baris data yang terduplikasi

ii. *Identify outlier and imputation the outlier*:
Terdapat 13 data pencilan dalam dataset




```
Outlier
    Math score    = [23, 13, 13]
    Writing score = [28, 29, 27, 28, 31, 29]
    Reading score = [24, 23, 26, 23]
```

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

Kemudian untuk 2 metode feature selection:

iii. RFE

iv. Chi-square

b. 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

a. Langkah-langkah yang saya lakukan dalam mengerjakan proyek ini:

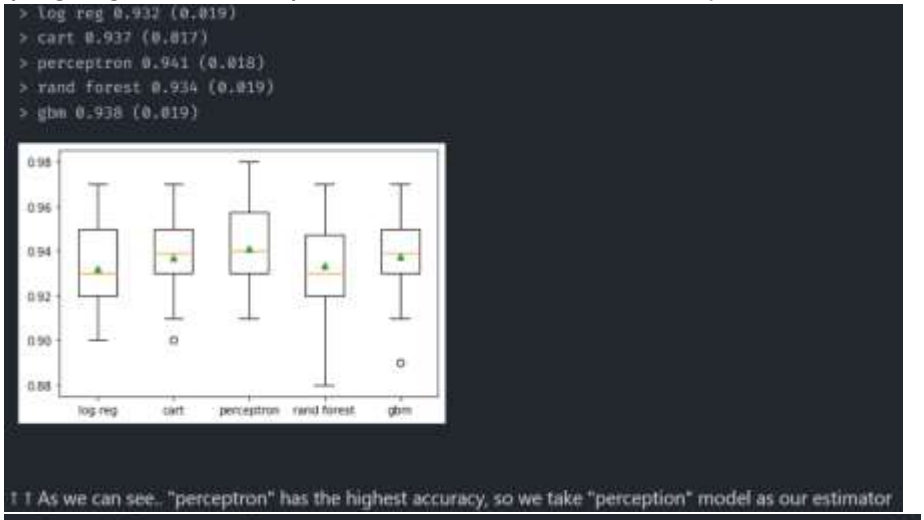
i. Import library yang dibutuhkan

ii. Load and check dataset

iii. Data cleaning → 1. Delete duplicated rows → 2. Identify outlier & impute to it's 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 >= 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.




```
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_associate's degree	parental level of education_master's degree	parental level of education_some high school
0	1	0	0	0	0	0	0
1	0	0	1	0	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	1	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
   a. Dataset diperoleh dari kaggle.com https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams
   b. Cara memperoleh dataset: mengunduh dari website
   c. Terdapat 2 metode *cleaning data*: Delete duplicated rows, Identify outlier & Imputation.
   d. Terdapat 2 metode feature selection: RFE, Chi-square
   e. Pendekatan analisis statistika
   f. Hasil observasi dataset:
      i. Students with a higher parental's degrees have more chances to achieve a higher scores
      ii. Based on gender, female have higher score average than male
      iii. The females dominate the higher scores in reading and writing
      iv. Male outperform female in ONLY math
      v. Along this dataset female and male students are equally distributed
      vi. "Lunch" is quite influential in performing student's exams score