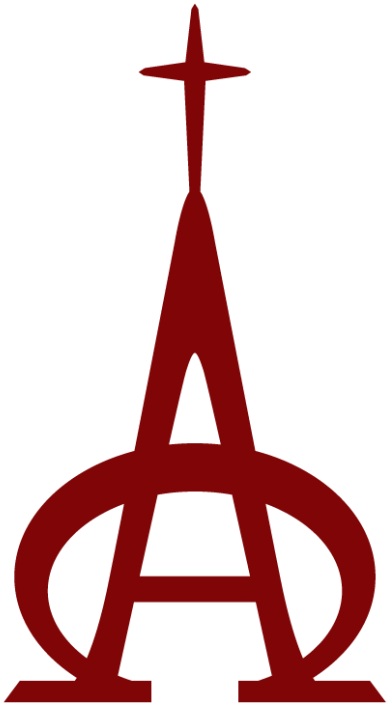
Laporan UAS IBDA 3111 Nomor 1



Ditulis oleh:

Moody Asyer - 191900154
Prodi: IBDA

CALVIN INSTITUTE OF TECHNOLOGY
JAKARTA

2021

Latar Belakang

Setiap tahun ada kurang lebih 10000 mahasiswa yang mendapatkan beasiswa LPDP untuk melanjutkan studi di Universitas terbaik di dalam maupun luar negeri. Untuk mendapatkan beasiswa, mahasiswa harus lolos berbagai seleksi dan tes yang diadakan.

Diketahui bahwa rektorat CIT mengadakan suatu survei yang hasilnya terdapat kurang lebih 50% mahasiswa IBDA yang berencana untuk studi lanjut dan beasiswa LPDP menjadi pilihan yang tepat. Oleh karena itu, yang perlu dilakukan adalah mempelajari bagaimana pola penerimaan beasiswa LPDP terkhususnya di seleksi akademik. Karena itu, dalam kasus ini kita akan membuat model pembelajaran mesin untuk mendapatkan akurasi setinggi-tingginya untuk dapat mempelajari penerimaan beasiswa LPDP.

Paparan Studi Kasus

Pada kasus kali data yang akan digunakan adalah data dari dataset 'Hasil Tes SBK LPDP Gelombang I - 2021.csv' nantinya pada dataset tersebut akan dilakukan teknik prapemrosesan dan evaluasi model pembelajaran mesin.

Analisa

Keterangan:

Penanda untuk menghitung jumlah teknik pra pemrosesan data yang digunakan

Model ke-	Metode pengolahan data yang dilakukan	Algoritma yang digunakan	Akurasi model
1	1. Mencari dan mengisi *missing value.* 2. Menghapus data duplikat. 3. Mencari *single value.* 4. Mencari *low value.* 5. Meng-*encode* data kategorik dengan one hot encoder. 6. Melakukan *scaling* dengan robust scaler pada data numerik. 7. Mentransformasi data numerik agar lebih gaussian dengan metode *uniform quantile transform.* 8. Meng-*encode* data output yang bersifat kategorik dengan *label encoder.*	Logistic Regression	89.89110%
2	1. Mencari dan mengisi *missing value.* 2. Menghapus data duplikat. 3. Mencari *single value.* 4. Mencari *low value.* 5. Meng-*encode* data kategorik dengan one hot encoder. 6. Melakukan *scaling* dengan Standard scaler pada data numerik	Random Forest Classifier	87.20930%

	7. Mentransformasi data numerik agar lebih gaussian dengan metode *normal quantile transform.* 8. Meng-*encode* data output yang bersifat kategorik dengan *label encoder*		
3	1. Mencari dan mengisi *missing value* *2.* Menghapus data duplikat 3. Mencari *single value.* 4. Mencari *low value.* 5. Meng-*encode* data kategorik dengan one hot encoder. 6. Melakukan *scaling* dengan Robust scaler pada data numerik 7. Mentransformasi data numerik agar lebih gaussian dengan metode *uniform quantile transform.* 8. Melakukan transformasi polynomial pada data 9. Meng-*encode* data output yang bersifat kategorik dengan *label encoder*	AdaBoost Classifier	86.58546%
4	1. Mencari dan mengisi *missing value* *2.* Menghapus data duplikat 3. Mencari *single value.* 4. Mencari *low value.* 5. Meng-*encode* data kategorik dengan one hot encoder. 6. Melakukan *scaling* dengan MinMaxScaler pada data numerik 7. Mentransformasi data numerik agar lebih gaussian dengan metode *yeo-johnson.* *8.* Meng-*encode* data output yang bersifat kategorik dengan *label encoder.*	Decision Tree Classifier	85.08859%
5	1. Mencari dan mengisi *missing value* *2.* Menghapus data duplikat 3. Mencari *single value.* 4. Mencari *low value.* 5. Meng-*encode* data kategorik dengan ordinal encoder. 6. Melakukan *scaling* dengan RobustScaler pada data numerik 7. Mentransformasi data numerik agar lebih gaussian dengan metode *uniform quantile transform.* *8.* Meng-*encode* data output yang bersifat kategorik dengan *label encoder.*	SVM	80.39867%

- Dari kelima model yang dibuat, didapatkan bahwa model ke 1 yang menggunakan algoritma Logistic Regression memiliki akurasi tertinggi, yaitu 89.89110%. Sisanya

memiliki akurasi yang kurang lebih sama, karena jumlah teknik pra pemrosesannya sama hanya berbeda metode saja.

- Pada tahap pra pemrosesan data, missing value tidak saya hapus karena satu data tersebut tidak akan dianggap kesalahan, tetapi lebih dianggap bahwa seorang mahasiswa tersebut belum mengisi kampus tujuannya.
- Untuk data duplikat akan dihapus karena dirasa sangat tidak mungkin untuk semua kategori bahkan hingga hasil tes memiliki data yang sama persis. Hal ini akan dianggap suatu kesalahan sehingga yang duplikat akan dihapus.
- Untuk pencarian low value sebenarnya ditemukan, namun saya memilih untuk tidak menghapusnya karena semua data yang ada dianggap asli.
- Selain itu, sebenarnya pada data juga memiliki hal-hal lain yang ditemukan seperti low variance, outliers, dsb (tidak dicantumkan), namun hal ini akan saya sikapi sama dengan menganggap bahwa setiap data tersebut asli sehingga tidak boleh dihapus.
- Seleksi fitur juga tidak dilakukan karena semua fitur yang ada akan diperhitungkan penting/berpengaruh pada output, sehingga kita tidak akan menghapus fitur yang ada.
- Untuk setiap model, dapat dilihat bahwa encoder yang digunakan akan sangat berpengaruh besar kepada akurasi, hal ini bisa dilihat untuk model 1-4 dimana encoder untuk data input yang digunakan adalah one hot encoder yang mengubah setiap datanya menjadi unik (matriks) dan keakuratannya tetap terjaga, sedangkan untuk model 5 encoder yang digunakan adalah ordinal encoder yang mengubah setiap datanya menjadi angka urut (seperti ada ranking). Perbedaan keunikan ini dapat dilihat akan sangat mempengaruhi akurasinya. Kita bisa melihat bahwa ketika menggunakan ordinal encoder akurasinya akan turun hingga kurang lebih hampir berbeda 10%.
- Untuk setiap model dibanding model ke 1 mengalami penurunan sedikit, hal ini bisa saja terjadi karena metode yang lain baik itu scaling dan normalisasi/standarisasi terhadap memiliki pengaruh yang berbeda kepada data dan ditemukan bahwa transformasi uniform quantile transform (membuat data lebih gaussian) dan scaling dengan robust scaler memiliki pengaruh yang lebih baik kepada data sehingga bisa didapatkan akurasi yang lebih tinggi (meningkatkan hasil akurasi dari hasil pelatihan model pembelajaran mesin).
- Pada kasus ini khususnya pada model ke 3, dilakukan transformasi polynomial yang mengubah fitur dari dataset. Fungsinya adalah agar pelatihan lebih mudah dilakukan, tapi benar saja ada dampak berupa turunnya akurasi yang didapatkan karena transformasi ini.
- Setiap fitur pada kasus ini kita anggap penting, fitur atau faktor yang paling berpengaruh pada kasus kelulusan di seleksi akademik bagi calon awardee sebenarnya bisa kita lihat ada beberapa hal:
    - Yang pertama tentunya merupakan data numerik yang merupakan jumlah jawaban benar pada tes.
    - Tetapi, data kategorikal khususnya seperti jenis beasiswa, negara tujuan, menggunakan LoA atau tidak juga menjadi faktor penting yang berpengaruh pada hasil diterima atau ditolak mahasiswa.
    - Yang terakhir adalah bidang studi yang dipilih dan kampus tujuan, terlepas dari seputar pembahasan ini kita tahu bahwa setiap bidang studi dan kampus

memiliki standar/nilai penerimaan seorang mahasiswa yang berbeda-beda, sehingga faktor ini juga mempengaruhi hasil apakah mahasiswa lolos atau tidak.