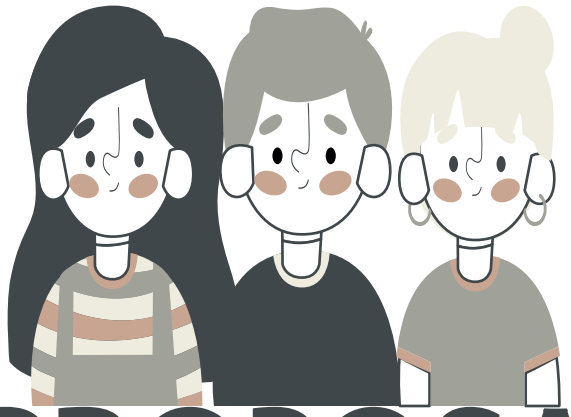


PROPOSAL

ANALISIS KETEPATAN MAHASISWA SARJANA LULUS 4 TAHUN

by: Dhea Fajriati Anas

Practice Case 1

Introduction to Data Science

*Nama instansi, data, dan hal-hal di dalam proposal ini hanya bersifat fiktif dan digunakan sebagai latihan

OUTLINE

PROLOGUE

Jurusan Ekonomi Pembangunan di Universitas Satu Nusa mempunyai tujuan untuk terakreditasi Internasional. Salah satu syarat untuk terakreditasi Internasional adalah 70% mahasiswanya harus lulus tepat waktu yaitu, 4 tahun.

Namun, kenyataannya tidak sampai 50% mahasiswa yang lulus tepat waktu hingga tahun ajaran 2020/2021. Oleh karena itu, pihak jurusan ingin mengetahui faktor-faktor apa saja yang mempengaruhi tingkat kelulusan mahasiswa, sehingga dapat diatasi sejak dini.

BUSSINESS UNDERSTANDING

Problem Statement: tidak ada suatu sistem yang dapat mengetahui berapa banyak mahasiswa yang berpotensi tidak lulus tepat waktu dan lulus tepat waktu setiap semester

Objectives: membuat sistem dapat mengetahui berapa banyak mahasiswa yang berpotensi tidak lulus tepat waktu setiap dan lulus tepat waktu setiap semester.

Expected Output: pada aplikasi web kemahasiswaan diberikan visualisasi terkait persen dan jumlah mahasiswa yang berpotensi tidak lulus tepat waktu setiap dan lulus tepat waktu setiap semester.

DATA UNDERSTANDING

Technique: data mahasiswa diambil dari database kemahasiswaan jurusan ekonomi pembangunan pada tahun 2010-2020.

Adapun atribut data yang digunakan adalah

1. Tahun dan bulan masuk
2. Tahun dan bulan lulus
3. Nama
4. Nomor Induk Mahasiswa
5. Tempat/Tanggal Lahir
6. Asal SMA
7. Penerima beasiswa/tidak
8. Sedang Bekerja/Tidak
9. Rata-rata Penghasilan Orang Tua
10. Status Menikah/belum
11. Indek Prestasi Semester 1-8
12. IPK
13. Lulus tepat waktu/tidak

DATA PREPARATION

Data Pre-Processing yang akan dilakukan adalah

1. Mengecek apakah data yang duplicate, dan menghapusnya
2. Mengecek apakah ada data yang missing values dan menindaklanjutinya
3. Menambah kolom baru berisi label berdasarkan kolom Lulus tepat waktu/tidak menjadi 1(tepat waktu) dan 0 (tidak tepat waktu)
4. Melakukan hal yang sama seperti poin nomor 3 untuk kolom Penerima beasiswa/tidak, Sedang Bekerja/tidak, dan Sudah menikah/belum
5. Mengecek tipe data setiap kolom apakah sudah sesuai, jika belum disesuaikan

MODELLING

Modelling: algoritma yang digunakan adalah klasifikasi Random Forest. Random Forest adalah kumpulan dari decision tree yang beroperasi menjadi suatu gabungan fungsional (Renata dan Ayub 2020). Random Forest dipilih karena kesalahan prediksi pada satu decision tree dapat ditutupi dengan kebenaran yang didapatkan dari decision tree lainnya yang benar, berjalan efisien pada data yang jumlahnya banyak, dan dapat berjalan baik dengan kelas yang populasinya tidak seimbang (Renata dan Ayub 2020).

Untuk mengatasi data kelas yang tidak seimbang digunakan metode *Synthetic Minority Oversampling Technique* (SMOTE) over-sampling. Pendekatan ini melakukan over-sampling pada kelas minoritas.

EVALUATION

Evaluation: metode yang digunakan adalah confusion matrix. Confusion matrix adalah suatu alat yang memiliki fungsi untuk melakukan analisis apakah classifier tersebut baik dalam mengenali tuple dari kelas yang berbeda. Confusion matrix merupakan matriks yang menampilkan prediksi klasifikasi dan klasifikasi yang aktual (Mahardhika et al. 2015).

Penghitungan kinerja model dapat dilakukan dengan menghitung nilai accuracy, precision, recall, dan F1-Score. *Accuracy* menggambarkan seberapa akurat model dalam mengklasifikasikan dengan benar. *Precision* menggambarkan akurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model. *Recall* atau *sensitivity* menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. F-1 Score menggambarkan perbandingan rata-rata precision dan recall yang dibobotkan

EVALUATION

		Actual Values	
		True	**False**
Predicted	**True**	TP Correct Result	TN Unexpected Result
	False	FN Missing Result	FN Correct Absence of Result

Gambar 1. Confusion matrix menampilkan total positive dan negative tuple

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1Score = \frac{2xRecallxPrecision}{Recall + Precision}$$

Keterangan:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

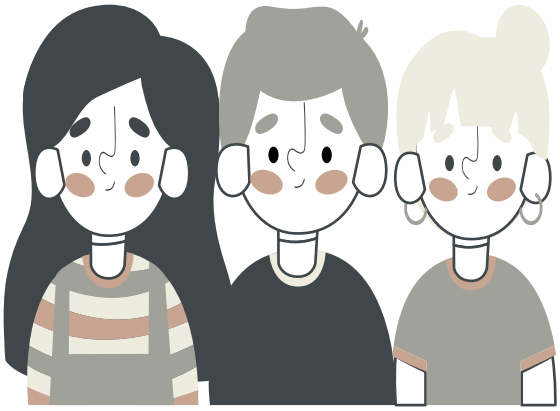
PROJECT TIMELINE AND BUDGETING

No	Kegiatan	Mei				Juni				Juli			
		1	2	3	4	1	2	3	4	1	2	3	4
1	Membuat arsitektur data												
2	Mengumpulkan data												
3	Data Pre-Processing												
4	Labeling												
5	Feature Extraction												
9	Modelling												
10	Evaluation												

Data Scientist

Data Engineer

Detail	Unit Price (Rupiah)	Duration (hari)	Total Price (Rupiah)
Data Science	500.000/hari	30	15.000.000
Data Engineer	700.000/hari	60	42.000.000



THANK YOU !