LEMBAR KERJA MAHASISWA (LKM)

LK.8 Perancangan Project Data Science

Nama	**: Agung Adi Rangga**
Tanggal	**: 12/4/2025**
Kelas	**: 5A**
Judul Project	**:** Prediksi Risiko Kegagalan Siswa dengan membandingan 5 model dan pemilihan Model Terbaik

A. Instruksi

Peserta diminta untuk merancang sebuah proyek Data Science yang berfokus pada permasalahan di bidang pendidikan. Rancangan proyek ini harus disusun secara sistematis berdasarkan metodologi CRISP-DM (Cross Industry Standard Process for Data Mining) yang mencakup enam tahapan utama, yaitu:

1. Business Understanding (Pemahaman Bisnis)
2. Data Understanding (Pemahaman Data)
3. Data Preparation (Persiapan Data)
4. Modeling (Pemodelan)
5. Evaluation (Evaluasi)
6. Deployment (Penerapan)

Pada setiap tahapan, peserta diharapkan dapat:

1. Menjelaskan tujuan dan fokus kegiatan pada tahap tersebut.
2. Menguraikan langkah-langkah yang dilakukan serta teknik atau metode yang digunakan.
3. Menjelaskan jenis dan sumber data yang diperlukan.
4. Menunjukkan hasil atau keluaran yang diharapkan dari tiap tahap.

Gunakan contoh kasus nyata atau permasalahan aktual di dunia pendidikan, seperti: Prediksi prestasi belajar siswa, Analisis tingkat kehadiran, Deteksi dini siswa berisiko tidak lulus, atau Rekomendasi pembelajaran adaptif berbasis data.

Hasil akhir dari tugas ini berupa dokumen rancangan proyek Data Science lengkap yang menggambarkan alur proses dari awal hingga implementasi model, serta menunjukkan bagaimana solusi berbasis data dapat memberikan manfaat nyata bagi peningkatan mutu pendidikan.

B. Format Perancangan

Tahapan CRISP-DM	Instruksi untuk Peserta	Rancangan Implementasi
1. Business Understanding (Pemahaman Bisnis)	1. Pilih konteks pendidikan (contoh: sekolah, universitas, pelatihan). 2. Identifikasi permasalahan yang dapat diselesaikan dengan data science. 3. Rumuskan tujuan bisnis (contoh: meningkatkan prestasi siswa, menurunkan tingkat ketidakhadiran).	1. Sekolah menengah atas di lingkungan perkotaan dengan latar belakang sosial-ekonomi beragam. 2. Permasalahan DataScience a. Tingginya risiko kegagalan akademik karena keterbatasan dukungan belajar dan ketidakhadiran. b. Intervensi belum terpersonalisasi

		sehingga sumber daya tidak efisien. c. Kesetaraan layanan belum terpantau sehingga potensi bias gender atau wilayah muncul. 3. Tujuan Bisnis a. Meningkatkan prestasi siswa dengan deteksi dini risiko kegagalan (target AUC-ROC ≥ 0.85). b. Menurunkan tingkat ketidakhadiran siswa melalui rekomendasi intervensi berbasis kausal (target efektivitas > 20%). c. Menjamin fairness layanan dengan gap metric < 5% untuk gender dan alamat.
2. Data Understanding (Pemahaman Data)	1. Jelaskan sumber data (contoh: data nilai siswa, absensi, data keluarga). 2. Sebutkan jenis data (numerik, kategorikal, teks, waktu). 3. Deskripsikan fitur dan target yang akan digunakan.	1. Sumber Data ● Data Nilai Siswa (G1, G2, G3) rekam nilai periode untuk pelacakan perkembangan. ● Data Absensi (absences, ketidakhadiran) frekuensi tidak hadir, dapat berasal dari sistem presensi sekolah. ● Data Keluarga & Dukungan (parents' education, duplas dukungan, internet, status keluarga) membantu memahami konteks sosial. ● Data Demografi & Aktivitas (usia, jenis kelamin, alamat, aktivitas ekstra) memberikan profil siswa. 2. Jenis Data a. Numerik: nilai, absensi, studytime,

		failures, health, ketidakhadiran, progression rate. b. Kategorikal: sekolah, jenis_kelamin, alamat, dukungan_sekolah, dukungan_keluarga, internet_rumah, status_ortu. c. Teks: alasan_sekolah, pekerjaan orang tua (bisa dikodekan). 3. Fitur & Target a. Fitur utama: rata-rata nilai (G1, G2), total dukungan, absensi, studytime, failures, lifestyle_score, family_stability, binary dukungan. b. Fitur tambahan: encoding demografi (jenis_kelamin, alamat), target encoding untuk pekerjaan orang tua, engineering progression rate, flags (high_absences, low_studytime). c. Target: risiko_gagal (1 jika G3 < 10), ditargetkan untuk prediksi klasifikasi dengan fairness monitoring.
3. Data Preparation (Persiapan Data)	1. Tuliskan langkah pembersihan data: hapus duplikat, tangani nilai kosong, dan outlier. 2. Transformasi data: normalisasi, encoding data kategorikal.	1. Langkah Pembersihan Data ● Hapus duplikat: df.drop_duplicates(inplace=True) untuk pastikan setiap observasi unik. ● Tangani nilai kosong: identifikasi dengan df.isnull().sum(), lalu isi dengan strategi

		(median untuk numerik, modus untuk kategorikal) atau hapus baris jika proporsi kecil. ● Deteksi outlier: gunakan IQR (Q1-1.5*IQR, Q3+1.5*IQR) untuk masing-masing fitur numerik dan filter baris ekstrem yang tidak representatif. 2. Transformasi Data 2.1. Normalisasi numerik: terapkan StandardScaler atau MinMaxScaler agar fitur berbagi skala yang sama sebelum pemodelan. 2.2. Encoding kategorikal: ● Binary (yes/no) → astype(int) ● Label encoding untuk ordinal ● Target encoding atau one-hot untuk fitur nominal penting 2.3. Feature engineering tambahan: buat indikator (misal high_absences, low_studytime) dan agregasi dukungan keluarga untuk memperkaya sinyal.
4. Modeling (Pemodelan)	1. Pilih algoritma yang sesuai (contoh: Decision Tree, Random Forest, Logistic Regression). 2. Jelaskan alasan pemilihan algoritma.	Saya menggunakan 5 algoritma untuk perbandingan dan mencari model terbaik (RF, LR, DT, XGBoost, Dan LightBGM) Algoritma yang paling sesuai: Regresi Linear Alasan: stabil terhadap noise, menangani fitur numerik/kategorikal setelah EDA campuran tanpa banyak tuning,

		memberikan feature importance untuk interpretasi, dan memiliki mekanisme ensembel yang membuatnya lebih tahan terhadap overfitting serta bisa ditingkatkan fairness dengan class weight.
5. Evaluation (Evaluasi)	Pilih metode evaluasi yang akan digunakan misalkan menggunakan cross-validation atau confusion matrix.	Confusion Matriks
6. Deployment (Penerapan / Implementasi)	Buat rancangan deploymentnya tampilan interface nya	1. Arsitektur Deployment ○ Model & pipeline dijalankan di container Python (Streamlit + FastAPI) dengan dependency dari requirements.txt. ○ Model best-predictive dan artefak scaler/encoder disimpan di /models. ○ API endpoint /predict untuk inference real-time, /explain untuk SHAP. ○ Background job (FastAPI + Celery) memproses batch prediction dari Dataset/TesPrediksi .csv. ○ Monitoring via Prometheus + Grafana (AUC, latency, fairness gap). ○ CI/CD: GitHub Actions menjalankan unit test + snyk_code_scan, kemudian deploy ke Azure App Service. 2. Tampilan Interface (Streamlit) ○ Header: Judul project, badge status Snyk, tanggal update. ○ Sidebar:

		■ Pilih mode (Predict, Explain, Dashboard). ■ Input siswa (dropdown sekolah, slider umur, checkboxes dukungan, absensi). ■ Tombol "Prediksi Risiko" & "SHAP Explain". ○ Main Area: ■ Predict mode: ■ Ringkasan skor (AUC, Fairness gap, Calibration). ■ Kartu hasil prediksi (risiko_gagal %, rekomendasi intervensi). ■ Explain mode: ■ SHAP force plot + summary bar chart. ■ Tabel top fitur + kontribusi. ■ Dashboard mode: ■ Grafis distribusi target, fairness per sensitif atribut, ROC curve. ■ Timeline simulasi policy

		(high-benefit vs high-risk).