

LEMBAR KERJA MAHASISWA (LKM)

LK.8 Perancangan Project Data Science

Nama	: NUR FAHIRA NURDIN
Tanggal	: 06 – Desember -2025
Kelas	: 5AI-A (Applied Machine Learning)
Judul Project	: Model Prediksi Risiko Putus Sekolah Tingkat Provinsi Berbasis Indikator Sumber Daya Sekolah Menggunakan Machine Learning

1. 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.

2. Format Perancangan

Tahapan CRISP-DM	Instruksi untuk Peserta	Rancangan Implementasi
1. Business Understanding (Pemahaman Bisnis)	<ol style="list-style-type: none"> Pilih konteks pendidikan (contoh: sekolah, universitas, pelatihan). Identifikasi permasalahan yang dapat diselesaikan dengan data science. Rumuskan tujuan bisnis (contoh: meningkatkan prestasi siswa, menurunkan Rancangan Implementasi tingkat ketidakhadiran). 	<ol style="list-style-type: none"> Konteks: Ketimpangan kualitas pendidikan antar provinsi di Indonesia. Masalah: Sulitnya menentukan prioritas alokasi bantuan sumber daya (guru/ruang kelas) untuk menekan angka putus sekolah. Tujuan: Membangun model prediksi untuk mengestimasi <i>dropout rate</i> berdasarkan rasio fasilitas sekolah, sehingga pemangku kebijakan dapat melakukan intervensi dini
2. Data Understanding (Pemahaman Data)	<ol style="list-style-type: none"> Jelaskan sumber data (contoh: data nilai siswa, absensi, data keluarga). Sebutkan jenis data (numerik, kategorikal, teks, waktu). Deskripsikan fitur dan target yang akan digunakan. 	<ol style="list-style-type: none"> Sumber Data: Statistik Sekolah 2024/2025 dari Kemendikdasmen (File: dataset_sekolah.xlsx). Jenis Data: Tabular (Numerik & Kategorikal). Fitur (X): Jumlah Siswa, Guru, Ruang Kelas, Siswa Mengulang, Status Sekolah (Negeri/Swasta). Target (Y): Angka Putus Sekolah (<i>Dropout</i>) .
3. Data Preparation (Persiapan Data)	<ol style="list-style-type: none"> Tuliskan langkah pembersihan data: hapus 	<ol style="list-style-type: none"> Data Cleaning: Menghapus prefiks "Prov." pada nama

	<p>duplikat, tangani nilai kosong, dan outlier.</p> <p>2. Transformasi data: normalisasi, encoding data kategorikal.</p>	<p>provinsi dan konversi tipe data <i>string</i> ke <i>numeric</i>.</p> <p>2. Feature Engineering: Membuat variabel baru berupa rasio (<i>Student_Teacher_Ratio</i>, <i>Class_Crowdedness</i>, <i>Repeater_Rate</i>) agar data antar provinsi yang besar dan kecil bisa dibandingkan secara adil.</p> <p>3. Encoding: Mengubah "Negeri/Swasta" menjadi angka 0/1 .</p>
4. Modeling (Pemodelan)	<p>1. Pilih algoritma yang sesuai (contoh: Decision Tree, Random Forest, Logistic Regression).</p> <p>2. Jelaskan alasan pemilihan algoritma.</p>	<p>1. Menggunakan dua algoritma regresi: Decision Tree (sebagai baseline yang mudah diinterpretasi) dan Random Forest (untuk akurasi lebih tinggi melalui <i>ensemble learning</i>).</p> <p>2. Membagi data menjadi 80% Training dan 20% Testing.</p> <p>3. Melakukan pelatihan model untuk mempelajari pola hubungan antara rasio fasilitas dan tingkat putus sekolah .</p>
5. Evaluation (Evaluasi)	Pilih metode evaluasi yang akan digunakan misalkan menggunakan cross-validation atau confusion matrix	<p>1. Menggunakan K-Fold Cross-Validation (k=5) untuk menguji kestabilan model pada data yang sedikit.</p> <p>2. Metrik Evaluasi: MAE (Mean Absolute Error) untuk melihat rata-rata kesalahan prediksi dalam persen, dan R-Squared.</p> <p>3. Analisis <i>Feature Importance</i> untuk mengetahui faktor dominan penyebab putus sekolah .</p>
6. Deployment (Penerapan / Implementasi)	Buat rancangan deploymentnya tampilan interface nya	Implementasi model menggunakan framework Gradio berbasis web. Pengguna dapat

		memasukkan data sekolah (jumlah siswa, guru, dll) dan sistem akan menampilkan prediksi risiko putus sekolah beserta rekomendasi kebijakan (Aman/Waspada/Bahaya) secara <i>real-time</i> .
--	--	---