LEMBAR KERJA MAHASISWA (LKM)

LK.8 Perancangan Project Data Science

Nama	**: RAMADHINI ANJANI HAMID**
Tanggal	**: 03 Desember 2025**
Kelas	**: 5AI-B**
Judul Project	**: Prediksi Jumlah Siswa Berdasarkan Jenis Kelamin dan Status Sekolah**

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. Konteks pendidikan yang dipilih adalah **data jumlah siswa Sekolah Dasar (SD) pada tingkat Kabupaten/Kota tahun 2025**. Dataset ini berisi informasi jumlah siswa berdasarkan kategori sekolah (Negeri dan Swasta) serta jenis kelamin (Laki-laki dan Perempuan).

		Pemilihan konteks ini relevan karena data pendidikan merupakan salah satu indikator penting dalam perencanaan kebutuhan sekolah dan analisis kualitas pendidikan di suatu wilayah. 2. Permasalahan yang dapat dipecahkan melalui data science adalah **estimasi atau prediksi jumlah siswa pada tahun berikutnya** berdasarkan pola jumlah siswa laki-laki dan perempuan, baik di sekolah negeri maupun swasta. Dengan pemodelan regresi, data science dapat membantu pemerintah daerah atau dinas pendidikan mengantisipasi kebutuhan fasilitas sekolah, guru, dan sarana pendukung lainnya. 3. Tujuan bisnis dari proyek ini adalah **membangun model prediksi jumlah siswa SD** agar instansi pendidikan dapat melakukan perencanaan yang lebih akurat, seperti penyediaan ruang kelas, tenaga pendidik, dan alokasi anggaran. Selain itu, model ini dapat membantu dalam memonitor tren perkembangan jumlah siswa serta mendukung pengambilan keputusan berbasis data.
2. Data Understanding (Pemahaman Data)	1. Jelaskan sumber data (contoh: dataa nilai siswa, absensi, data keluarga). 2. Sebutkan jenis data (numerik, kategorikal, teks, waktu). 3. Deskripsikan fitur dan target yang akan digunkan.	1. Sumber data yang digunakan berasal dari dataset **"Jumlah Siswa SD Tahun 2025"** yang berisi informasi jumlah siswa berdasarkan kategori sekolah (Negeri dan Swasta) serta jenis

		kelamin (Laki-laki dan Perempuan) pada tingkat Kabupaten/Kota. Data ini merupakan data tabular yang diambil dalam bentuk file **CSV** dan diunggah ke Google Colab untuk proses analisis. 2. Jenis data numerik dan kategorikal. 3. Fitur yang digunakan dalam proyek ini terdiri dari jumlah siswa berdasarkan kategori jenjang dan status sekolah, yaitu *Laki-laki Negeri*, *Laki-laki Swasta*, *Perempuan Negeri*, dan *Perempuan Swasta*. Keempat fitur ini dipilih karena mewakili komposisi utama populasi siswa pada setiap kabupaten/kota. Sementara itu, target yang digunakan adalah kolom *Total*, yaitu jumlah keseluruhan siswa pada wilayah tersebut. Target ini menjadi variabel yang diprediksi oleh model untuk mengetahui estimasi total jumlah siswa berdasarkan distribusi kategori siswa yang tersedia.
3. Data Preparation (Persiapan Data)	1. Tuliskan langkah pembersihan data: hapus duplikat, tangani nilai kosong, dan outlier. 2. Transformasi data: normalisasi, encoding data kategorikal.	Pada tahap persiapan data, dilakukan beberapa proses pembersihan untuk memastikan dataset siap digunakan dalam pemodelan. Pertama, data diperiksa dan dihapus dari duplikasi agar tidak terjadi pengulangan informasi yang dapat mempengaruhi hasil analisis. Selanjutnya, dicek nilai kosong dan dipastikan tidak ada data yang hilang sehingga tidak mengganggu proses perhitungan model. Karena seluruh kolom berupa numerik, tidak diperlukan

		proses encoding data kategorikal dan tidak dilakukan normalisasi karena algoritma regresi linear masih dapat bekerja efektif tanpa transformasi tersebut. Dataset kemudian ditambahkan kolom *Total* sebagai nilai target untuk diprediksi oleh model.
4. Modeling (Pemodelan)	1. Pilih algoritma yang sesuai (contoh: Decision Tree, Random Forest, Logistic Regression). 2. Jelaskan alasan pemilihan algoritma.	Tahap pemodelan dilakukan dengan memilih algoritma **Linear Regression** karena sesuai untuk kasus prediksi nilai numerik berbasis hubungan linear antar fitur. Algoritma ini dipilih karena sederhana, mudah diinterpretasikan, dan mampu memberikan performa yang sangat baik pada dataset dengan pola hubungan yang jelas antar variabel. Model dilatih menggunakan data training untuk mempelajari hubungan antara jumlah siswa berdasarkan kategorinya dengan total siswa secara keseluruhan. Setelah model terbentuk, model diuji untuk memastikan bahwa performanya stabil dan dapat digunakan dalam proses prediksi.
5. Evaluation (Evaluasi)	Pilih metode evaluasi yang akan digunakan misalkan menggunakan cross-validation atau confusion matrix	Model diuji menggunakan data testing dan dievaluasi menggunakan MAE, RMSE, dan R^2 Score. Nilai evaluasi yang diperoleh yaitu MAE = 45.66, RMSE = 48.96, dan R^2 = 0.99999. Hasil ini menunjukkan model memiliki tingkat akurasi yang sangat tinggi dan mampu memprediksi total siswa hampir tepat dengan data asli.
6. Deployment (Penerapan/Implementasi)	Buat rancangan deploymentnya tampilan interface nya	Model disimpan dalam format .pkl menggunakan pickle. Selanjutnya model di-deploy menggunakan **Gradio** agar bisa digunakan melalui antarmuka web sederhana. Pengguna cukup memasukkan

	jumlah siswa laki-laki dan perempuan baik di sekolah negeri maupun swasta, lalu sistem menampilkan hasil prediksi total siswa secara otomatis.