PROJECT PLAN
Dicoding Bootcamp - Capstone Project

ID Tim Capstone Project : DB9-G007

Judul Proyek **:** SmartSpend

Tema yang dipilih **: Akses Keuangan untuk Semua**

Learning Path : Data Science

List Anggota **:**

1. B25B9D044 - Sisilia Manullang- Data Science - **Aktif**
2. B25B9D043 - I Gusti Putu Ardan Setiawan - Data Science - **Aktif**
3. B25B9D042 - Wendy Shabirin Kadarsyah- Data Science - **Aktif**
4. B25B9D040 - Dimas Ahda Sabila - Data Science - **Aktif**

A. Ringkasan Eksekutif

Pengelolaan keuangan pribadi tetap menjadi tantangan bagi banyak individu, meskipun layanan digital berkembang pesat. Informasi transaksi yang tersebar di berbagai kanal dan tidak tersusun rapi menyebabkan keputusan finansial sering diambil tanpa dasar data yang kuat. Minimnya waktu, alat, dan kemampuan untuk memahami pola pengeluaran secara komprehensif semakin memperbesar kesenjangan ini. Di sisi lain, proses analisis keuangan pribadi yang tidak efisien dan tidak konsisten membuat pengguna sulit mengidentifikasi kategori pengeluaran dominan maupun potensi pemborosan.

Proyek ini dibangun untuk menjawab tantangan tersebut melalui pendekatan berbasis data dan *machine learning*. Pertanyaan kunci yang mendorong pengembangan meliputi bagaimana transaksi dapat dikelompokkan secara akurat, sejauh mana algoritma mampu mengungkap pola perilaku finansial pengguna, serta bagaimana sistem dapat menyederhanakan analisis keuangan sehingga pengguna dapat memahami kondisi finansial mereka secara cepat dan jelas. Proyek in menghadirkan solusi *end-to-end* yang

mengotomatisasi ekstraksi data dari struk belanja, mengkategorikan transaksi secara cerdas, dan menghasilkan insight finansial yang dapat ditindaklanjuti.

Output utama proyek ini berupa dashboard analitik interaktif yang menyajikan ringkasan transaksi, distribusi pengeluaran per kategori, tren pengeluaran bulanan, skor kesehatan finansial, profil perilaku pengguna, dan rekomendasi personal yang dirancang untuk mendorong pengelolaan keuangan yang lebih baik. Proyek ini dipilih karena relevansinya dengan kebutuhan pasar serta kontribusinya dalam meningkatkan literasi finansial melalui solusi digital yang praktis dan mudah diakses. Dengan kombinasi analitik komprehensif dan visualisasi yang intuitif, membantu pengguna membuat keputusan finansial yang lebih terinformasi, terukur, dan selaras dengan tujuan keuangan jangka panjang mereka.

B. Cakupan Proyek dan Hasil Kerja

Proyek ini berfokus pada pengembangan sebuah platform analitik keuangan pribadi yang memanfaatkan teknologi ***Optical Character Recognition (OCR)*** dan ***machine learning*** untuk menyederhanakan proses pencatatan dan analisis transaksi keuangan. Ruang lingkup proyek dirancang secara strategis agar dapat diselesaikan dalam periode **4–5 minggu**, dengan prioritas menghasilkan prototipe yang memiliki fitur Inti, yang memberikan nilai tertinggi bagi pengguna, serta batasan yang jelas untuk menjaga ketepatan waktu dan kualitas penyampaian.

1. Fitur Utama

berikut adalah kemampuan inti yang akan dikembangkan dalam prototipe. Tabel ini menunjukkan ruang lingkup fitur *must-have* untuk MVP, beserta batasan fungsionalnya:

Komponen	Deskripsi Fungsi	Prioritas	Cakupan Implementasi
OCR Extraction	Mengubah gambar struk/screenshot menjadi teks mentah, lalu mengekstraksi	Tinggi	OCR berbasis library (Tesseract/EasyOCR), regulasi pola untuk ekstraksi data

	tanggal, nominal, merchant, dan deskripsi		
Data Cleaning & Normalization	Membersihkan noise OCR, menstandarkan format angka, tanggal, dan merchant	Tinggi	*Cleaning* pipeline, normalisasi teks, validasi tipe data
Classification Model	Mengkategorikan transaksi ke dalam kelas pengeluaran (makanan, transportasi, belanja, dll.)	Tinggi	Model *ML supervised* (*Random Forest/Logistic Regression*), evaluasi dasar
Clustering Model	Mengelompokkan pengguna berdasarkan perilaku finansial	Tinggi	K-Means atau algoritma sejenis, profil cluster sederhana
Financial Health Scoring	Memberikan penilaian kondisi finansial 0–100	Tinggi	Formula + model regresi ringan, indikator saving/spending
Analytics Dashboard	Visualisasi transaksi, kategori, tren bulanan, dan profil perilaku	Tinggi	Dashboard Streamlit dengan grafik interaktif
Recommendation Engine	Memberikan rekomendasi perbaikan finansial	Rendah / *Nice to Have*	*Rule-based* + *insight* dari pola *cluster*

Tabel 1 Fitur Utama

2. Pembagian Tanggung Jawab Tim

Pembagian tanggung jawab akan dirancang secara kolaboratif, dimana setiap modul/fitur akan dikerjakan oleh minimal 2 orang untuk mengurangi beban teknis dan menurunkan resiko bottleneck. Setiap anggota akan tetap memiliki fokus tanggung jawab utama untuk menjaga arah kerja, namun proses pengerjaan akan dilakukan secara lintas-modul dengan pembagian tugas yang lebih fleksibel.

Modul	Penanggung Jawab Utama	Tugas Inti	Kolaborator
OCR & Data Extraction	Anggota A	- Implementasi OCR dasar (gambar → teks) - Ekstraksi tanggal, nominal, merchant - Penyusunan pola teks (regex sederhana)	- Anggota B - Anggota C - Anggota D
Data Cleaning & Feature Engineering	Anggota B	- Cleaning hasil OCR - Normalisasi format angka/tanggal - Penyusunan fitur dasar untuk ML	- Anggota A - Anggota C - Anggota D
Machine Learning	Anggota C	- Model klasifikasi kategori - *Clustering* perilaku - *Health scoring* model	- Anggota A - Anggota B - Anggota D
Dashboard & Integration	Anggota D	- UI/UX Streamlit - Integrasi OCR → Processing → ML - Visualisasi data & rekomendasi - *Deployment*	- Anggota A - Anggota B - Anggota C

Tabel 2. Pembagian Tanggung Jawab

C. Jadwal Pengerjaan

Pengerjaan project akan akan direncanakan lewat penjadwalan berdasarkan pada fitur-fitur yang ada pada Tabel 1. Fitur Utama.

1. Kegiatan Mingguan

Minggu	Fokus utama	Kegiatan	hasil
Minggu 1	*Foundation*, Data *Exploration* & OCR *Research*	- Pengumpulan data (sintetik) berdasarkan e.g. sample daftar penjualan brand FMCG di MT dan GT - Setup repo, workspace, dan environment - Pengumpulan 20–30 sampel struk/screenshot - Riset *library* OCR (Tesseract/EasyOCR) - Pemetaan pola transaksi (tanggal, nominal, merchant) - EDA awal terhadap sampel	- Arsitektur sistem awal - Keputusan library OCR - Dataset sampel awal - Dokumen pola transaksi

		- Kickoff meeting + pembagian peran	
	OCR pipeline & Data *Processing Development*	- Implementasi OCR pipeline (A+B) - *Cleaning & normalisasi* hasil ekstraksi (B+C) - Penyusunan regex sederhana - Penyusunan dataset versi 1 - Validasi internal hasil ekstraksi - Review integrasi awal dengan tim dashboard	- `ocr_extractor.py` - `cleaner.py` - Dataset v1 (siap untuk ML dasar)
Minggu 2	*Classification Model*	- Kumpulkan dan siapkan data latih berlabel. - Ekstraksi fitur (misal TF-IDF pada nama merchant, binning jumlah, atribut waktu). - Pilih algoritma (misal *Random Forest*) dan latih model, *tuning* parameter. - Evaluasi model (akurasi ≥80%, *precision/recall* ≥70%). - Simpan model terlatih ke file.	Model klasifikasi transaksi terlatih tersimpan (misal `transaction_classifier.pkl`) dan kemampuan menambahkan kolom kategori prediksi pada data transaksi.
	Clustering	Buat fitur tingkat-user dengan mengagregasi data transaksi (misal distribusi kategori, rata-rata transaksi, volatilitas pengeluaran). - Terapkan algoritma cluster (misal K-Means) dan tentukan jumlah cluster optimal (elbow method, Silhouette Score). - Evaluasi kualitas klaster. - Tetapkan label klaster ke setiap user.	Setiap pengguna berlabel klaster tertentu, beserta profil klaster (persona) dan visualisasi klaster.
Minggu 3	*Financial Health Scoring*	- Definisikan komponen skor (misal ***saving rate***, **adherence anggaran**, **stabilitas belanja**, dll) dan bobotnya (domain *knowledge*). - Buat skema skor sintetis sebagai *ground truth*. - Latih model regresi	Skor kesehatan finansial per user (0–100) dengan label kategorinya, serta breakdown komponen skor.

		(misal *Random Forest*) untuk prediksi skor (target 0–100). - Evaluasi performa. - Buat fungsi interpretasi skor (misal kategori *Poor/Fair/Good/Excellent*).	
	Analytics Dashboard	Rancang dan bangun UI dashboard menggunakan Streamlit. - Kembangkan visualisasi (grafik pie kategori, tren pengeluaran bulanan, KPI card, dsb). - Integrasi output model *(cluster, health score*, rekomendasi) ke dalam *dashboard*. - Deploy aplikasi ke cloud (*Streamlit Community Cloud* atau sejenis).	Dashboard interaktif online yang menampilkan ringkasan transaksi, pola pengeluaran, skor finansial, dan rekomendasi secara visual.
Minggu 4	*Recommendation Engine*	- Definisikan logika rekomendasi berbasis aturan (menggunakan klaster pengguna, skor kesehatan, pola belanja, dll). - Hasilkan daftar rekomendasi personalisasi (minimal 5 saran relevan per user). - Prioritaskan rekomendasi berdasarkan dampak (tinggi/sedang/rendah). - Integrasi rekomendasi ke dashboard (tampilan dalam halaman khusus).	Daftar rekomendasi finansial personal untuk setiap pengguna (5–7 rekomendasi prioritas dengan langkah aksi jelas).

Tabel 3. Kegiatan Mingguan

Rencana Pengerjaan timeline dapat dilihat pada Google sheet berikut ini:
Timeline_Project_Capstone

D. Sumber Daya Proyek

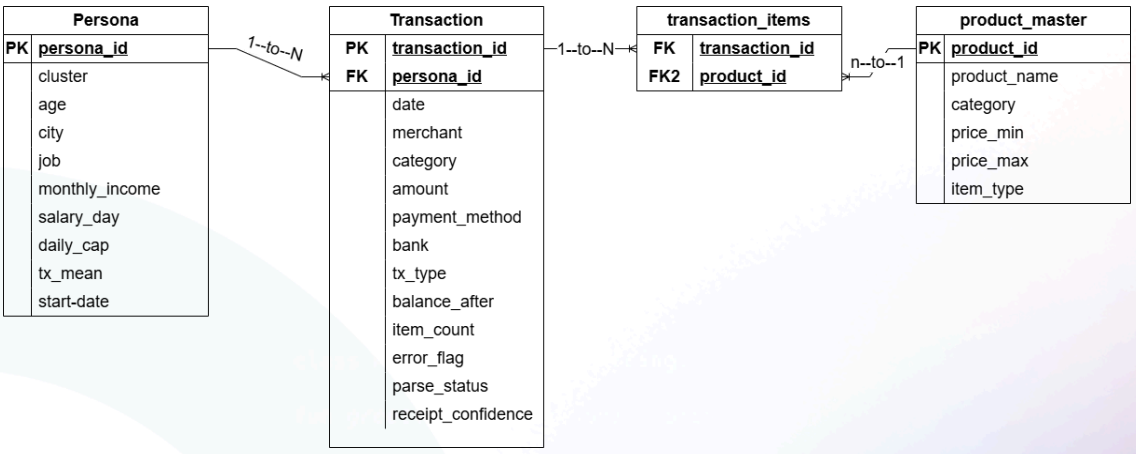
Sumber daya atau *resource* yang diperlukan dalam pengerjaan proyek: Bahasa Pemrograman,*Framework*, API, cloud backend, dataset, paper/journals/articles, serta sumber daya lainnya yang diperlukan. **Jelaskan fungsi dari setiap**

sumber daya/tools yang akan digunakan secara singkat namun jelas fungsinya.

1. **Bahasa Pemrograman**
   Python akan menjadi bahasa pemrograman utama untuk pekerjaan Capstone Project.

2. **Informasi Dataset**
   Dataset ini menyajikan gambaran komprehensif mengenai perilaku transaksi pengguna, pola pengeluaran, dan aktivitas finansial yang realistis berdasarkan konteks pengguna Indonesia. Dataset dirancang untuk mendukung berbagai analisis, termasuk:
   - Klasifikasi kategori transaksi
   - Segmentasi pengguna (persona modelling)
   - Analisis perilaku keuangan
   - Ekstraksi data dari struk (OCR parsing)
   - Pembuatan dashboard analitik
   - Deteksi anomali dan anomali spending

   Dataset mencakup 100 pengguna (persona) dengan lebih dari 100.000 transaksi, lengkap dengan detail merchant, item-level, *subscription*, dan teks struk yang mensimulasikan hasil OCR. Setiap entri memberikan wawasan mendalam terhadap kebiasaan finansial individu, memungkinkan eksplorasi menyeluruh dalam membangun sistem analisis keuangan modern dan model prediktif.



Gambar 1. ERD Dataset

3. **Cloud Backend**
   Tim akan menggunakan github untuk melakukan kolaborasi pengerjaan Capstone *Project* ini.

4. **Paper/Journals/Articles**
   Dengan menggunakan referensi, dapat ditemukan algoritma paling tepat dalam melakukan *clustering* dan *classification* data untuk data pengelolaan keuangan pribadi. Melalui pemahaman teori yang mendalam, tim dapat berargumen mengapa algoritma dan model dipilih untuk mengolah data.

5. **Mentoring**
   Melalui masukan dan kritik yang didapat selama sesi evaluasi bersama mentor, tim dapat melakukan perbaikan untuk menyempurnakan proses dan hasil dari Capstone Project.

6. **Frontend**
   Output Capstone Project dapat diakses oleh user melalui ***Streamlit***

E. Rencana Manajemen Risiko dan Isu

Hasil identifikasi terhadap faktor-faktor yang dapat menjadi penyebab proyek gagal maupun tertunda. Tim dapat menggunakan analisis SWOT ataupun risk management framework untuk mengidentifikasi dan mencari solusi atas penyelesaian isu yang mungkin terjadi selama pengerjaan proyek ini.

Risiko	Kemungkinan	Dampak	Mitigasi	Kontingensi
Akurasi OCR (Pengenalan Teks) untuk Struk dengan bahasa Indonesia rendah	Tinggi	Tinggi	- Library OCR yang terbukti dengan dilatih data bahasa Indonesia -Standarisasi *pre-processing image* (contoh: crop, *grayscale*, *sharpening*) untuk meningkatkan akurasi input. - Sediakan	Batasi jenis struk yang didukung (contoh: hanya dari retailer besar yang formatnya konsisten).

			proses bisnis konfirmasi setelah struk dibaca apakah input sudah akurat atau tidak dan disediakan fitur mengedit inputasi struk	
Data sintetik tidak merepresentasikan realitas dengan baik	Tinggi	Tinggi	- Gunakan algoritma generative (contoh: GANs, Variational Autoencoders) yang canggih. - Perkaya prompt algoritma generatif dengan real data daftar produk dan harga dari pengumpulan manual di retailer besar	-Tim akan memperkaya prompt dengan melampirkan data real transaksi yang selama ini dilakukan di dunia nyata seperti transaksi online pribadi, pergi ke retail dan mencatat daftar produk, dll
Algoritma klasifikasi/clustering pengeluaran tidak akurat	Sedang	Tinggi	- Eksperimen dengan beberapa algoritma (contoh: *Random Forest*, K-Means). - Lakukan features engineering yang mendalam. - Gunakan teknik evaluasi model (*cross-validation*, confusion matrix) secara ketat.	Sederhanakan kategori pengeluaran atau gunakan model yang lebih interpretable seperti Decision Tree.
Metrik "Kesehatan Keuangan" yang dibuat subjektif	Sedang	Sedang	- Lakukan penelitian literatur tentang indikator kesehatan keuangan (cash	Perkaya prompt untuk algoritma generatif yang

atau tidak relevan			*flow*, rasio tabungan, *debt-to-income*) untuk berbagai kelas pendapatan di Indonesia - Validasi metrik dengan pembimbing capstone project	menghasilkan data sintetik dengan detail matriks keuangan dari masing-masing kelas pendapatan di Indonesia