Project Plan

Prediksi Risiko Penyakit Kronis Berbasis Data Sosial-Ekonomi dan Gaya Hidup

ID Grup : LAI25-SM082

Anggota Grup :

1. A001XBM122 – Dewi Kunthi Siswati Suryo – IPB University – [Aktif]
2. A001XBM196 – Halida Fiadnin – IPB University – [Aktif]
3. A001XBM223 – Illinia Malika Putri – IPB University – [Aktif]
4. A001YBM349 – Muhammad Rizky Fajar – IPB University – [Aktif]

Tema Pilihan:
Keberlanjutan dan Kesejahteraan

Nama proyek Capstone:
Prediksi Risiko Penyakit Kronis Berbasis Data Sosial-Ekonomi dan Gaya Hidup

Ringkasan Eksekutif/Abstrak:

Penyakit kronis seperti diabetes, hipertensi, dan penyakit jantung semakin banyak menyerang kelompok usia produktif hingga lanjut usia akibat perubahan pola hidup yang tidak sehat dan tekanan sosial-ekonomi. Padahal, risiko ini sebenarnya dapat ditekan melalui deteksi dini dan intervensi gaya hidup yang tepat sejak awal. Saat ini, belum banyak sistem prediksi yang secara komprehensif menyasar berbagai kelompok usia dengan mempertimbangkan faktor sosial-ekonomi dan gaya hidup sebagai indikator risiko. **Proyek ini bertujuan** untuk mengembangkan model machine learning berbasis data sosial-ekonomi dan kebiasaan hidup untuk memprediksi risiko penyakit kronis pada masyarakat umum, sehingga mereka dapat mengambil langkah pencegahan lebih awal.

Menurut Hindawi 2022, penyakit kronis semakin meningkatkan angka kematian manusia, dan pengobatan untuk penyakit ini menghabiskan lebih dari 70% pendapatan pasien, sehingga sangat penting untuk meminimalkan faktor risiko pasien yang dapat menyebabkan kematian. Oleh karena itu, **pernyataan masalah** yang ingin diatasi adalah

tingginya angka penyakit kronis akibat kurangnya alat prediksi yang mudah diakses dan berbasis data nyata, yang menciptakan kesenjangan antara kondisi saat ini dan upaya preventif yang seharusnya dilakukan. **Pertanyaan penelitian** yang diajukan adalah "Faktor apa yang paling signifikan dalam meningkatkan risiko penyakit kronis pada berbagai kelompok usia?" dan "Bagaimana performa berbagai algoritma machine learning dalam memprediksi risiko penyakit kronis berbasis data gaya hidup?" Proyek ini penting karena dapat mendorong kesadaran kesehatan preventif di masyarakat luas dan membantu mengurangi beban ekonomi jangka panjang akibat penyakit kronis.

Bagaimana grup Anda menghasilkan proyek ini?

Proyek ini dimulai dengan pemahaman mendalam tentang pentingnya deteksi dini dalam pencegahan penyakit kronis. Proyek ini berfokus pada pemecahan masalah yang berkaitan dengan tingginya angka penyakit kronis yang dipengaruhi oleh faktor sosial, ekonomi, dan pola hidup. Langkah pertama yang dilakukan adalah mengidentifikasi masalah utama, yaitu kurangnya sistem prediksi yang mengintegrasikan data sosial-ekonomi dan gaya hidup untuk kelompok usia rentan terhadap penyakit kronis. Proyek dimulai dengan pengumpulan data gaya hidup, riwayat medis, dan faktor sosial-ekonomi, kemudian dilanjutkan dengan *preprocessing* dan pengembangan model *machine learning*. Setelah evaluasi model, sistem prediksi diimplementasikan untuk membantu individu mengambil langkah pencegahan dini terhadap penyakit kronis.

Lingkup & Hasil Proyek:

1. Garis Besar Batasan Proyek:

Proyek ini memiliki ruang lingkup yang terfokus pada pembuatan sistem prediksi risiko penyakit kronis berbasis machine learning dalam waktu 4 minggu. Proyek akan dikerjakan menggunakan data terbuka (open-source) dan **tidak mencakup**:

- Pengumpulan data primer langsung dari pasien/partisipan.
- Validasi medis secara klinis terhadap hasil prediksi.
- Penggunaan sistem prediksi dalam praktik medis aktual.

Fokus utama proyek adalah membangun dan menyajikan prototipe berbasis Streamlit Web App yang dapat menerima input pengguna dan memberikan prediksi risiko untuk beberapa penyakit kronis secara bersamaan.

2. Pembagian Tugas Tim (Tim berjumlah 4 orang):

Nama Anggota	Peran Utama	Tanggung Jawab Spesifik
Halida Fiadnin	Data Engineer & Analyst	Pencarian dataset, eksplorasi awal, pembersihan data, visualisasi EDA
Dewi Kunthi Siswati Suryo	Machine Learning Developer	Model baseline, model multi-output neural network, tuning
Illinia Malika Putri	Evaluator & Interpretability Specialist	Evaluasi model, feature importance (SHAP), visualisasi hasil
Muhammad Rizky Fajar	App Developer & Dokumentator	Pembuatan aplikasi Streamlit, layout UI, dokumentasi teknis dan presentasi

3. Hasil Kerja Terukur (Deliverables):

- Dataset siap pakai (clean, encoded, normalised)
- Model baseline (Logistic Regression/Random Forest)
- Model final (Multi-output Neural Network)
- Evaluasi model (AUC/F1-score per label, F1-score minimum 85%)
- Visualisasi interpretatif (SHAP, grafik risiko)
- Aplikasi Streamlit interaktif
- Dokumentasi proyek (README, laporan akhir, presentasi)

4. Rencana Harian (Ringkasan Pembagian Tugas Harian selama 4 Minggu):

Minggu	Hari	Aktivitas Utama
Minggu 1	Hari 1-2	Kick-off meeting, studi literatur
	Hari 3-5	Pencarian dan eksplorasi dataset, pembersihan data awal
	Hari 6-7	EDA, pembuatan dokumentasi awal hasil eksplorasi
Minggu 2	Hari 8-10	Preprocessing lanjutan, feature engineering
	Hari 11-12	Modeling baseline: Logistic Regression, Random Forest
	Hari 13-14	Evaluasi model baseline, mulai perancangan arsitektur NN
Minggu 3	Hari 15-17	Implementasi multi-output neural network, tuning parameter
	Hari 18-19	Evaluasi model akhir, pembuatan visualisasi interpretasi (SHAP/feature importance)
	Hari 20-21	Integrasi awal model ke Streamlit
Minggu 4	Hari 22-24	Pengembangan layout UI/UX Streamlit, validasi fungsi aplikasi
	Hari 25-26	Simulasi penggunaan, debugging akhir
	Hari 27-28	Penyusunan laporan akhir serta pembuatan slide dan video

5. Indikator Keberhasilan Setiap Minggu:

- **Minggu 1:** Dataset siap pakai dan EDA selesai.
- **Minggu 2:** Model baseline selesai dan model multi-output mulai dibangun.
- **Minggu 3:** Model multi-output selesai dan performa baik.
- **Minggu 4:** Aplikasi Streamlit selesai dan dokumentasi lengkap.

Jadwal Proyek:

Milestone	Minggu 1	Minggu 2	Minggu 3	Minggu 4
Riset Teoritis				
Pengumpulan & Eksplorasi Data				
Preprocessing & Feature Engineering				
Baseline Modeling (LogReg, RF)				
Modeling Multi-Output Neural Network				
Evaluasi & Interpretasi Model				
Pengembangan Aplikasi Streamlit				
Pengujian Akhir & Simulasi				
Dokumentasi & README				
Pengumpulan Project				

No	Milestone	Output
1	Riset Teoritis	Studi literatur dan mempelajari teori penilaian risiko penyakit serta hal-hal yang mempengaruhi
2	Pencarian & persiapan dataset	Dataset siap pakai untuk modelling
3	Eksplorasi dan visualisasi data awal	Insight awal, missing value, distribusi, outlier

4	Preprocessing & Feature Engineering	Dataset bersih & siap modelling
5	Baseline modeling (LogReg, RF)	Model dasar untuk komparasi
6	Bangun model multi-output neural network	Model utama proyek
7	Evaluasi & interpretasi model	AUC, F1, SHAP, analisis feature importance
8	Integrasi ke Streamlit	Halaman input, prediksi, dan output visual
9	Pengujian akhir & dokumentasi	Prediksi berbagai skenario, readme, laporan akhir
10	Pengumpulan Project	Pengumpulan seluruh berkas laporan, video presentasi, dan hasil project

Berdasarkan pengetahuan grup Anda, alat/IDE/Perpustakaan dan sumber daya apa yang akan digunakan grup Anda untuk menyelesaikan masalah?

Alat, IDE, library, dan sumber daya yang digunakan adalah sebagai berikut.

1. IDE: Google Colab (utama), Jupyter Notebook, Visual Studio Code
2. Bahasa Pemrograman: Python
3. Library:
    - TensorFlow: Library utama untuk membangun dan melatih model.
    - TensorFlow Decision Forests (TF-DF): Untuk implementasi model tree-based seperti Random Forest atau Gradient Boosting dalam ekosistem TensorFlow.
    - Pandas: Untuk manipulasi dan pengolahan data tabular.
    - NumPy: Untuk operasi numerik yang efisien.
    - Scikit-learn: Untuk pembagian data (train-test split), evaluasi metrik, dan preprocessing tambahan.
    - Matplotlib/Seaborn: Untuk visualisasi data dan hasil model.
4. Sumber Data: Dataset publik dari Kaggle.
5. Perangkat: Laptop pribadi masing-masing anggota tim.

Berdasarkan pengetahuan dan eksplorasi Anda, untuk apa grup Anda membutuhkan dukungan?

Grup kami memerlukan akses ke dataset yang relevan, terutama yang memuat data gaya hidup dan faktor sosial-ekonomi yang bisa digunakan untuk pelatihan model. Kami juga membutuhkan bimbingan dari mentor yang memahami penerapan machine learning di bidang kesehatan, agar proses pemilihan fitur dan evaluasi model berjalan lebih terarah. Beberapa referensi literatur atau studi sebelumnya tentang prediksi penyakit kronis juga

akan kami manfaatkan sebagai acuan agar pengembangan model ini tetap realistis dan sesuai dengan konteks penelitian.

Berdasarkan pengetahuan dan eksplorasi Anda, jelaskan kepada kami bagian Machine Learning dari Capstone Anda!

Kami akan membangun model machine learning multi-output classification menggunakan TensorFlow dan TensorFlow Decision Forests untuk memprediksi beberapa penyakit kronis. Setiap penyakit dilatih dengan dataset publik dari Kaggle, dan hasil model digunakan untuk inferensi sederhana melalui input data pasien.

Berdasarkan perencanaan grup Anda, apakah ada potensi Risiko atau Masalah yang dapat diidentifikasi terkait dengan proyek Anda?

1. Keterbatasan data
    - Faktor risiko: Keterbatasan data yang tidak lengkap atau tidak representatif dapat mempengaruhi akurasi model prediksi.
    - Rencana penanganan: Data akan dikumpulkan dari berbagai sumber yang dapat memberikan gambaran lebih luas. Jika data yang tersedia tidak cukup, data *augmentation* akan digunakan.
    - Pengendalian: Pemeriksaan data yang digunakan secara rutin. Selain itu, penanganan data hilang seperti imputasi atau penghapusan data yang tidak lengkap akan diterapkan.
2. Bias dalam data
    - Faktor risiko: Jika data yang digunakan tidak seimbang atau mewakili kelompok tertentu secara berlebihan, model dapat mengalami bias yang mempengaruhi prediksi.
    - Rencana penanganan: Data akan dianalisis terlebih dahulu untuk mengidentifikasi ketidakseimbangan kelas atau representasi yang tidak merata di berbagai kelompok. Jika ditemukan ketidakseimbangan data, metode seperti *oversampling* atau *undersampling* akan diterapkan untuk memastikan distribusi data yang seimbang.
    - Pengendalian: Evaluasi akan dilakukan untuk memastikan bahwa model tidak terpengaruh oleh bias. Jika ditemukan bias, model akan diperbaiki dan hasil prediksi akan diuji untuk berbagai kelompok.
3. Kualitas model
    - Faktor risiko: Kualitas model yang buruk atau tidak teroptimasi dapat menyebabkan prediksi yang tidak akurat atau tidak dapat digeneralisasi dengan baik ke data yang lebih luas.

- Rencana penanganan: Menguji berbagai algoritma untuk mengevaluasi kinerja model dan menentukan model yang paling efektif. Selain itu, dilakukan *cross-validation* dan *hyperparameter tuning* untuk mengoptimalkan model.
- Pengendalian: Mengevaluasi kinerja model dengan menggunakan data uji yang terpisah untuk memastikan bahwa model tetap efektif dan dapat diterapkan dalam berbagai kondisi.

Catatan/keterangan lain yang perlu kami pertimbangkan tentang aplikasi grup Anda

Proyek ini memiliki potensi untuk diterapkan dalam berbagai konteks, baik untuk individu yang ingin memantau risiko kesehatan mereka secara pribadi maupun untuk lembaga medis yang ingin menyediakan alat bantu deteksi dini penyakit kronis. Pengembangan lebih lanjut dapat mencakup fitur monitoring risiko yang lebih lanjut berdasarkan pola perkembangan data pribadi. Pengawasan tenaga ahli dan ketersediaan data sangat menentukan keberhasilan proyek ini.