DOKUMEN PROYEK

12S3205 - PENAMBANGAN DATA

CLASSIFICATION OF EXPLORING MENTAL DATA HEALTH USING LOGISTIC REGRESSION

Disusun Oleh :

12S22030	**Bryan Evans Simamora**
12S22049	**Agnes Monica Sanjani Harefa**
12S22050	**Yohana Christine Sitanggang**

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO (FITE)

INSTITUT TEKNOLOGI DEL

DAFTAR ISI

BAB 1

BUSINESS UNDERSTANDING

1.1 Determine Business Objective

Kesehatan mental merupakan salah satu aspek penting dalam kesejahteraan manusia yang sering kali terabaikan. Dalam dunia modern yang penuh tekanan, terutama di lingkungan kerja atau pendidikan, semakin banyak individu yang mengalami gangguan mental seperti depresi. Namun, banyak kasus depresi tidak terdeteksi secara dini, sehingga memperburuk kondisi penderita.
Melalui proyek ini, kami bertujuan untuk menganalisis data survei kesehatan mental guna mengidentifikasi faktor-faktor risiko yang berkontribusi terhadap depresi. Harapannya, hasil dari proyek ini dapat menjadi dasar untuk melakukan pencegahan dan memberikan dukungan yang tepat kepada individu yang rentan mengalami gangguan mental.

1.2 Determine Project Goal

- Tujuan Bisnis
- Mengidentifikasi faktor-faktor yang memengaruhi risiko gangguan mental di kalangan pekerja.
- Memberikan insight bagi perusahaan untuk membangun lingkungan kerja yang lebih sehat.
- Membantu pihak HR dan kebijakan dalam menyusun program dukungan mental health berdasarkan data.
- Menyediakan visualisasi yang membantu pemangku kebijakan memahami kondisi karyawan.

- Tujuan *Data Science*
- Membangun model klasifikasi untuk memprediksi apakah seseorang berisiko mengalami depresi.
- Melakukan EDA (*Exploratory Data Analysis*) untuk melihat distribusi data dan pola tersembunyi.
- Menghasilkan visualisasi untuk membantu stakeholder memahami temuan.

1.3 Produce Project Plan

- Kriteria Kesuksesan (*Success Criteria*)

Dari sisi teknis:
Model klasifikasi memiliki:
- Akurasi > 60%
- Precision > 60%
- Recall > 60
- Model berhasil di-deploy secara online dan dapat menerima input dari user.

Dari sisi bisnis:
- Model dapat digunakan sebagai alat bantu untuk deteksi dini risiko depresi.
- Hasil analisis dapat meningkatkan kesadaran dan perhatian terhadap isu kesehatan mental di lingkungan kerja maupun pendidikan.
- Model memberikan insight fitur mana yang paling memengaruhi risiko depresi.

- Kendala dan Asumsi

- Data yang digunakan bersifat sintetis, bukan data real-world, sehingga interpretasi hasil harus dilakukan dengan hati-hati dan tidak digunakan untuk diagnosis medis.
- Beberapa fitur mungkin memiliki nilai kosong (*missing value*) atau tidak konsisten, sehingga perlu dilakukan pembersihan dan pra-pemrosesan.
- Dataset berasal dari *survei global* → mungkin tidak sepenuhnya mewakili konteks lokal (seperti Indonesia), namun tetap relevan untuk studi dan pengembangan model.
- Model tidak ditujukan sebagai alat diagnosa medis, melainkan hanya sebagai tools prediksi dan analisis awal.

BAB 2

DATA UNDERSTANDING

2.1 Pengumpulan Data

Dataset yang digunakan dalam proyek ini bersumber dari Kaggle Playground Series S4E11. Dataset ini bersifat sintetis, dirancang untuk kebutuhan pembelajaran dan eksperimen dalam pengembangan model prediksi kesehatan mental.

- Format Data: CSV (Comma Separated Values)
- Jumlah Data:
  - Train Set: 140.700 baris, 20 kolom
  - Test Set: 60.300 baris (tanpa label Depression)
- Target Prediksi: Depression (0 = tidak depresi, 1 = depresi)

2.2 Describe Data

Data terdiri dari gabungan variabel kategorikal dan numerik, yang mencakup karakteristik demografis, kebiasaan harian, tekanan akademik/kerja, hingga aspek psikologis. Berikut beberapa fitur penting:

1. train.csv

- **Jumlah data**: 140.700 baris, 20 kolom
- **Kolom-kolom penting**:
  - Age: Rata-rata usia 40.4 tahun (18–60 tahun)
  - Gender: 2 kategori → Male, Female
  - City: 98 kota berbeda
  - Profession: 64 jenis pekerjaan, dengan Teacher paling banyak (24.906)
  - Academic Pressure, CGPA, Study Satisfaction: hanya terisi sekitar 20% data
  - Depression: 0 dan 1, dengan hanya **~18%** kasus depresi (klasifikasi imbalanced)

2. test.csv

- **Jumlah data**: 93.800 baris, 19 kolom (tanpa Depression)
- Distribusi mirip dengan train.csv
- **Nilai missing** di kolom seperti:
  - Academic Pressure, CGPA, Profession → banyak kosong
  - Gender, City, Degree dan lainnya → cukup lengkap

3. sample_submission.csv

- **Jumlah data**: 93.800 baris, 2 kolom:
  - id: Identik dengan test set
  - Depression: Default-nya semua bernilai 0 (placeholder untuk prediksi)

2.3 Validation Data

- Missing Value: Telah diidentifikasi dan akan ditangani dengan strategi seperti imputasi atau penghapusan.
- Outlier & Anomali: Nilai Age dalam rentang wajar (18–60), namun beberapa kategori seperti Sleep Duration dan Gender memiliki nilai non-standar.
- Data Duplikat: Belum ditemukan duplikasi signifikan pada kolom ID (unik).
- Distribusi Target: Dataset imbalanced → hanya sekitar 18% responden mengalami depresi (Depression = 1).