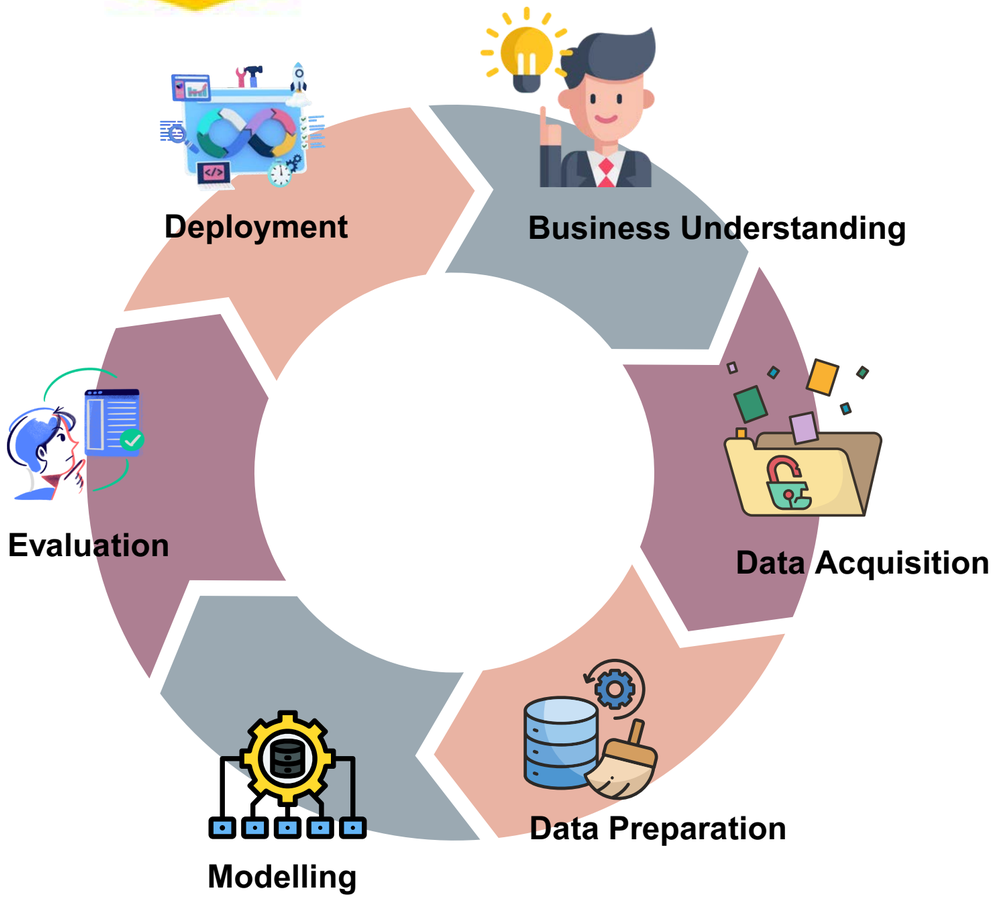
Pengantar Data Sains

#2 Meeting

Tahapan Proyek Data Sains

Ferdian Bangkit Wijaya, S.Stat., M.Si
NIP. 19900520202406100 1

Lifecycle Project



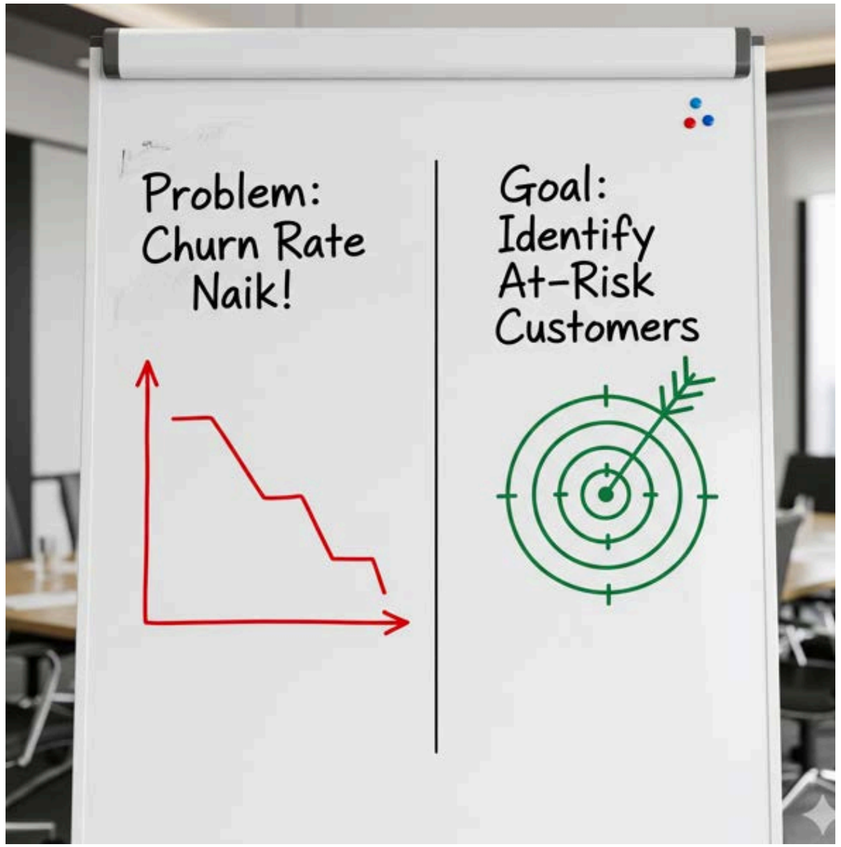
Model CRISP-DM (Cross-Industry Standard Process for Data Mining)

- Proyek data sains bisa menjadi kompleks dan penuh ketidakpastian.
- Tanpa kerangka kerja, proyek berisiko:
- Gagal menjawab pertanyaan bisnis yang tepat.
- Menghabiskan waktu pada data yang tidak relevan.
- Menghasilkan model yang tidak dapat digunakan di dunia nyata.
- Tidak memiliki metrik keberhasilan yang jelas.
- Kerangka kerja membantu kita tetap di jalur yang benar.

fokus pada tujuan bisnis (teknologi hanya alat, tujuan utamanya adalah menyelesaikan masalah nyata)

Chapman, P., et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc

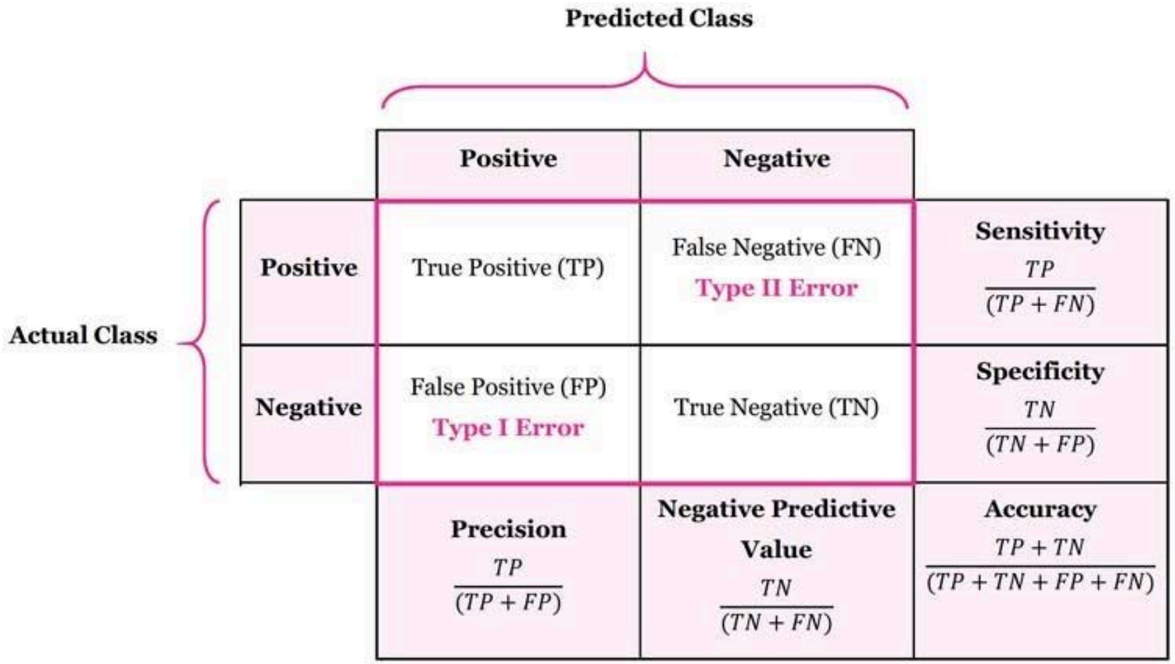
Business Understanding



- Studi Kasus: StreamFlix
- Situasi Bisnis: Perusahaan melihat adanya peningkatan jumlah pelanggan yang berhenti berlangganan (churn) dalam 3 kuartal terakhir. Biaya untuk mengakuisisi pelanggan baru 5x lebih mahal daripada mempertahankan yang sudah ada.
- Tujuan Bisnis: Mengurangi tingkat churn bulanan.
- Tujuan Proyek Data Sains: Membangun sebuah sistem untuk mengidentifikasi pelanggan yang berisiko tinggi untuk churn dalam 30 hari ke depan, agar tim marketing bisa melakukan intervensi.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.

Business Understanding

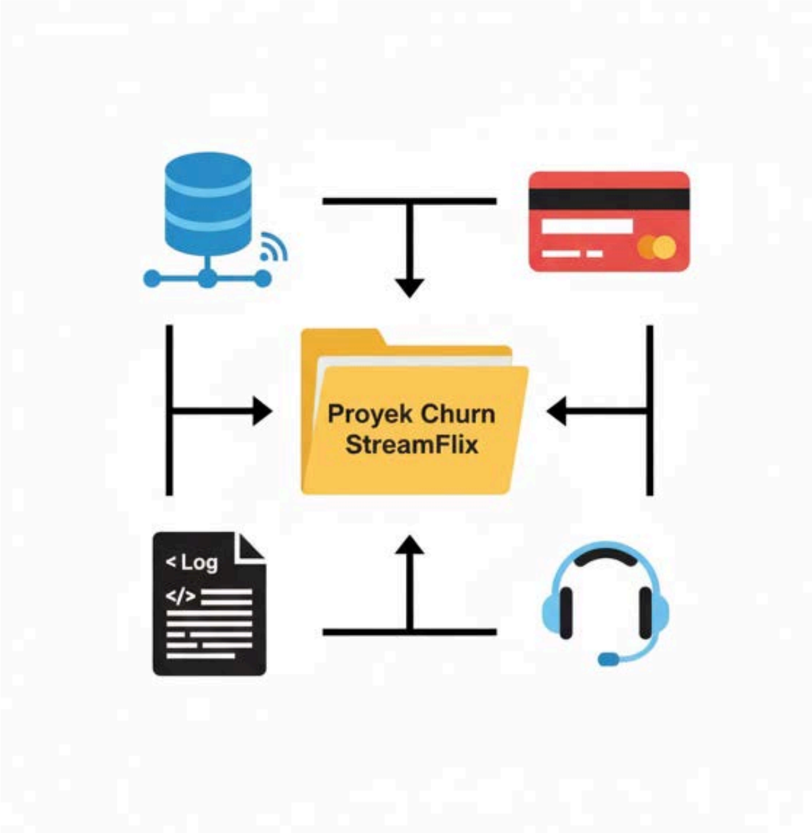


- Bagaimana kita tahu jika proyek ini berhasil?
- Kriteria Sukses Bisnis: "Mengurangi churn rate bulanan sebesar 2% dalam 6 bulan setelah model diimplementasikan."
- Kriteria Sukses Teknis: "Model harus mampu mengidentifikasi setidaknya 70% dari pelanggan yang sebenarnya akan churn (Metrik: Recall (Sensitivity) > 0.70)."

Recall penting di sini. "Bagi StreamFlix, lebih baik kita salah memberi penawaran diskon kepada pelanggan yang tidak akan churn (False Positive), daripada kita kehilangan pelanggan yang sebenarnya akan churn tapi tidak terdeteksi oleh model (False Negative). Jadi, 'menangkap' sebanyak mungkin pelanggan yang benar-benar akan churn adalah prioritas."

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.

Data Acquisition



- Mengumpulkan dan "berkenalan" dengan data yang tersedia.
- Sumber Data yang Dikumpulkan:
- Database Pelanggan (SQL): Data demografis (usia, lokasi), jenis langganan.
- Log Aktivitas Tontonan (File Log): Genre yang ditonton, jam tonton per hari, perangkat yang digunakan.
- Data Transaksi (SQL): Riwayat pembayaran, kegagalan pembayaran.
- Log Interaksi CS (CSV): Jumlah komplain, kategori masalah.

Data tidak tersimpan rapi di satu tempat. Tahap ini melibatkan "perburuan" data dari berbagai sistem dan departemen.

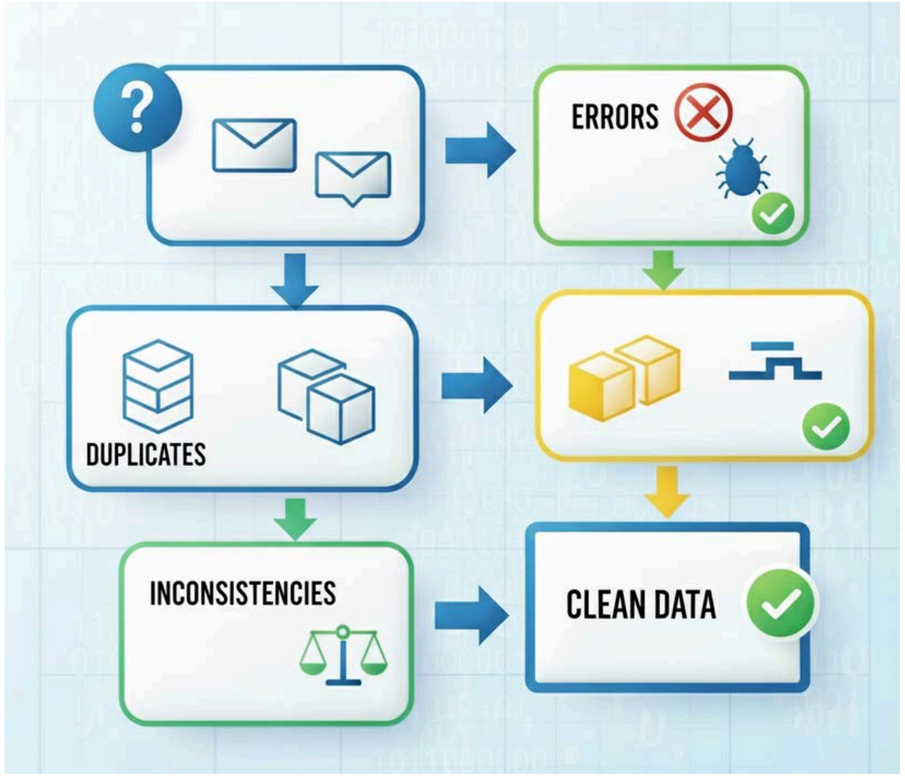
Data Preparation

- Tahap paling memakan waktu: Mengubah data mentah menjadi dataset bersih.
- "Garbage In, Garbage Out".
- Tugas Utama: Menggabungkan 4 sumber data yang berbeda (demografi, tontonan, transaksi, CS) menjadi satu tabel tunggal di mana setiap baris adalah satu pelanggan unik dan setiap kolom adalah fitur-fitur mereka.

Bayangkan, kita harus mencocokkan ID pelanggan dari sistem database, dengan ID pengguna dari file log, dan email dari log CS. Ini adalah pekerjaan rekayasa data yang rumit namun fundamental.

McKinney, W. (2017). Python for Data Analysis, 2nd Edition. O'Reilly Media.

Data Preparation



- Menangani nilai yang hilang, error, duplikat, dan inkonsistensi (Data Cleansing).
- Menangani Nilai Hilang: Kolom usia memiliki 10% data kosong. Tim memutuskan untuk mengisinya dengan nilai median usia pelanggan.
- Memperbaiki Error: Kolom perangkat berisi entri seperti "iphone", "iPhone 15", "Apple iPhone". Semua diseragamkan menjadi satu kategori: "iPhone".
- Menghapus Duplikat: Ditemukan beberapa pelanggan terdaftar dengan email yang sama, data duplikat dihapus.

McKinney, W. (2017). Python for Data Analysis, 2nd Edition. O'Reilly Media.

Data Preparation

- Menciptakan fitur baru yang lebih informatif dari data mentah (Feature Engineering)
- Dari log tontonan, dibuat fitur baru: rata_rata_jam_tonton_per_minggu, jumlah_genre_berbeda, rasio_film_selesai_ditonton.
- Dari log transaksi, dibuat fitur jumlah_kegagalan_pembayaran_3_bulan_terakhir.
- Dari log CS, dibuat fitur waktu_sejak_komplain_terakhir.
- Fitur-fitur ini jauh lebih kuat daripada data aslinya.

McKinney, W. (2017). Python for Data Analysis, 2nd Edition. O'Reilly Media.

Modeling

- Menggunakan statistik dan visualisasi untuk mendapatkan intuisi dari data (EDA).
- Visualisasi: Histogram menunjukkan pelanggan dengan jam tonton < 5 jam/minggu memiliki tingkat churn yang jauh lebih tinggi.
- Statistik: Rata-rata jumlah komplain ke CS untuk pelanggan yang churn adalah 3.2, sementara untuk yang tidak churn hanya 0.5.
- Hipotesis: Tingkat engagement (jam tonton) dan masalah teknis (komplain CS) adalah prediktor kuat untuk churn.

EDA membantu kita membentuk hipotesis. "Dari sini tim StreamFlix mulai curiga, jangan-jangan masalahnya ada di pengalaman menonton dan layanan pelanggan. Ini adalah petunjuk berharga untuk tahap selanjutnya."

Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

Modeling

- Menerapkan algoritma machine learning pada data yang telah disiapkan.
- Tipe Masalah: Klasifikasi Biner (Hasilnya hanya dua: Churn atau Tidak Churn).
- Pemilihan Algoritma: Tim memutuskan untuk mencoba 3 algoritma:
- Regresi Logistik (sebagai baseline sederhana).
- Random Forest.
- Gradient Boosting.
- Desain Uji: Data dibagi menjadi 80% data latih dan 20% data uji.

Beberapa model itu penting. "Kita tidak tahu pasti model mana yang terbaik. Jadi, kita latih beberapa model dan membandingkan performanya untuk memilih pemenangnya."

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

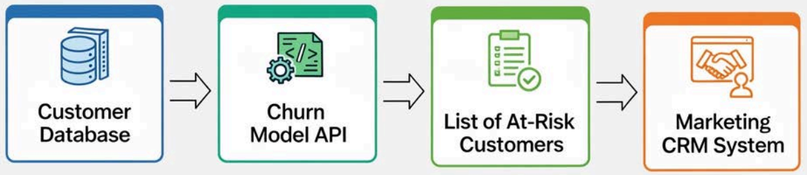
Evaluation

- Menilai apakah model memenuhi tujuan bisnis dan teknis.
- Evaluasi Model: Model Gradient Boosting menunjukkan performa terbaik di data uji.
- Hasil vs Kriteria Sukses:
- Performa Model: Recall = 78%.
- Kriteria Sukses Teknis: Recall > 70%.
- Keputusan: SUKSES! Model ini secara teknis memenuhi syarat untuk dilanjutkan ke tahap berikutnya.

Ini adalah momen "kelulusan" bagi model. "Dengan recall 78%, artinya dari 100 pelanggan yang benar-benar akan churn, model kita berhasil menangkap 78 di antaranya. Ini sudah melampaui target awal kita yaitu 70%. Tim sekarang yakin untuk merekomendasikan model ini kepada manajemen."

James, G., et al. (2013). An Introduction to Statistical Learning. Springer.

Deployment

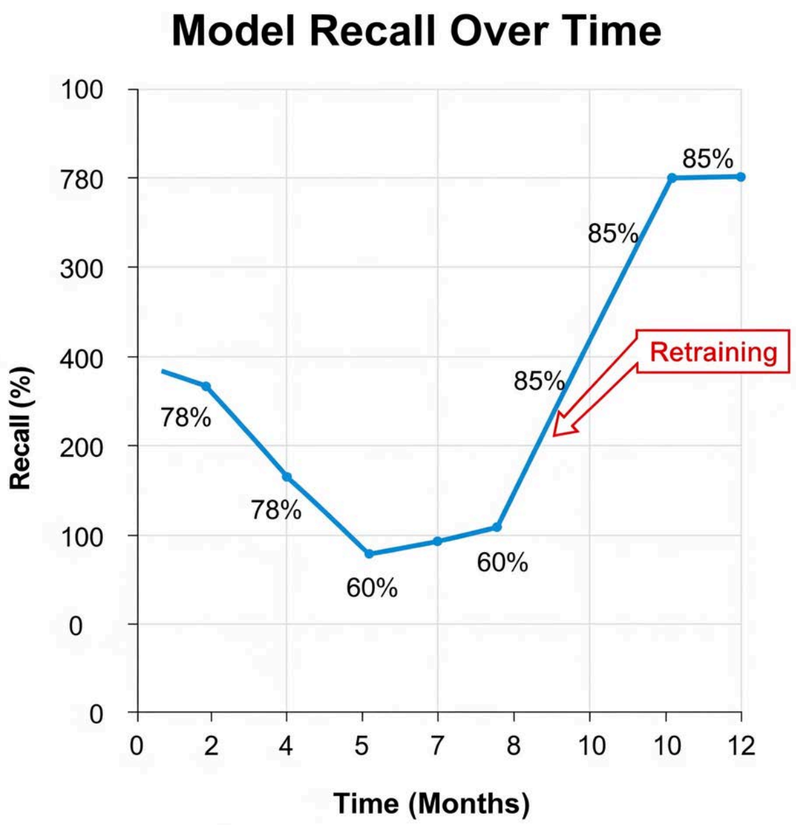


- Mengintegrasikan model ke dalam lingkungan produksi untuk memberikan nilai nyata (Integration)
- Rencana Deployment:
- Model di-hosting sebagai sebuah API di cloud.
- Setiap hari Senin, sebuah skrip otomatis akan berjalan, mengambil data pelanggan aktif.
- Skrip mengirim data ke API model untuk mendapatkan skor risiko churn untuk setiap pelanggan.
- Daftar 10% pelanggan dengan risiko tertinggi secara otomatis dikirim ke sistem CRM tim marketing.

Sekarang, setiap minggu, tim marketing tidak lagi menebak-nebak. Mereka mendapatkan daftar nama yang konkret, yang dihasilkan oleh data, untuk ditindaklanjuti dengan kampanye retensi (misalnya, email penawaran khusus)

Burkov, A. (2020). Machine Learning Engineering. True Positive Inc.

Deployment



- Pekerjaan tidak berhenti setelah deployment. Performa model harus terus dipantau (Monitoring).
- Monitoring: Sebuah dasbor dibuat untuk melacak metrik recall model setiap bulan.
- Model Drift: Setelah 6 bulan, tim melihat performa recall turun dari 78% menjadi 71% karena adanya perubahan perilaku pelanggan setelah rilis fitur baru.
- Pemeliharaan: Tim menjadwalkan model untuk dilatih ulang (retraining) setiap 3 bulan sekali menggunakan data baru agar performanya tetap optimal.

Dunia berubah, perilaku pelanggan berubah, produk pun berubah. Model kita harus ikut beradaptasi.

Burkov, A. (2020). Machine Learning Engineering. True Positive Inc.

Kerangka Kerja Alternatif: OSEMN

- OSEMN: Dibuat dari sudut pandang Data Scientist.
- Fokus pada alur kerja praktis sehari-hari.
- Lima Tahapan:
- Obtain (Memperoleh): Mengumpulkan data dari berbagai sumber (database, API, web scraping).
- Scrub (Membersihkan): Membersihkan, mem-format, dan memproses data mentah. Ini adalah data wrangling/munging.
- Explore (Mengeksplorasi): Melakukan EDA untuk menemukan pola, anomali, dan wawasan awal.
- Model (Memodelkan): Membangun model prediktif atau deskriptif.
- iNterpret (Menginterpretasikan): Menafsirkan hasil, menceritakan temuan (data storytelling), dan mengkomunikasikannya kepada pemangku kepentingan.

OSEMN lebih fokus pada siklus kerja individu atau tim kecil.

Mason, H., & Wiggins, C. (2010). A Taxonomy of Data Science. Data Gotham.

Kerangka Kerja Alternatif: SEMMA

SAS Institute Inc. Metodologi SEMMA.

- SEMMA: Dikembangkan oleh SAS Institute.
- Sangat fokus pada urutan tahapan teknis dalam pemodelan.
- Lima Tahapan:
- Sample (Mengambil Sampel): Memilih subset data yang representatif dari dataset yang sangat besar untuk dieksplorasi.
- Explore (Mengeksplorasi): Memvisualisasikan dan mencari hubungan atau anomali dalam data sampel.
- Modify (Memodifikasi): Mempersiapkan data, melakukan rekayasa fitur (feature engineering), dan transformasi.
- Model (Memodelkan): Menerapkan berbagai algoritma pemodelan untuk menemukan pola yang paling baik.
- Assess (Menilai): Mengevaluasi kegunaan dan keandalan model yang telah dibuat.

UNIVERSITAS
SULTAN AGENG
TIRTAYASA

UNTIRTA
Jawara
Jujur Adil Wibawa Amanah Religius Akuntabel

SEE YOU NEXT WEEK !

Ferdian Bangkit Wijaya, S.Stat., M.Si
NIP. 19900520202406100 1
ferdian.bangkit@untirta.ac.id