

**LAPORAN KERJA PRAKTIK**  
**DATA SCIENCE FOR BUSINESS DEVELOPMENT**  
**PENGUNAAN MODEL MACHINE LEARNING UNTUK**  
**MEMPREDIKSI *CHURN* NASABAH KARTU KREDIT PERBANKAN**  
**(STUDI KASUS DI PT COURSE-NET BANGUN INDONESIA)**

Diajukan untuk memenuhi persyaratan kelulusan

Mata Kuliah SIF339 - Kerja Praktek

Oleh :

**FAISAL AKBAR KUSPRIANTO / 302210009**



**PROGRAM STUDI SISTEM INFORMASI**  
**FAKULTAS TEKNOLOGI INFORMASI**  
**UNIVERSITAS BALE BANDUNG**  
**2025**

**LEMBAR PENGESAHAN**  
**PROGRAM STUDI SISTEM INFORMASI**  
**DATA SCIENCE FOR BUSINESS DEVELOPMENT**  
**PENGUNAAN MODEL MACHINE LEARNING UNTUK**  
**MEMREDIKSI *CHURN* NASABAH KARTU KREDIT PERBANKAN**  
**(STUDI KASUS DI PT COURSE-NET BANGUN INDONESIA)**

**Oleh :**

**Faisal Akbar Kusprianto / 302210009**

Disetujui dan disahkan sebagai

**LAPORAN KERJA PRAKTIK**

Bandung, Januari 2025

Koordinator Kerja Praktik Program Studi Sistem Informasi

**Rosmalina,ST.,M.Kom**

NIK.04104808122

**LEMBAR PENGESAHAN**  
**PEMBIMBING KERJA PRAKTIK**  
**DATA SCIENCE FOR BUSINESS DEVELOPMENT**  
**PENGUNAAN MODEL MACHINE LEARNING UNTUK**  
**MEMPREDIKSI *CHURN* NASABAH KARTU KREDIT PERBANKAN**  
**(STUDI KASUS DI PT COURSE-NET BANGUN INDONESIA)**

**Oleh :**

**Faisal Akbar Kusprianto / 302210009**

Disetujui dan disahkan sebagai

**LAPORAN KERJA PRAKTIK**

Bandung, Januari 2025  
Pembimbing Kerja Praktek

**Rosmalina,ST.,M.Kom**

NIK.04104808122

**LEMBAR PENGESAHAN**  
**PT COURSE-NET BANGUN INDONESIA**  
**DATA SCIENCE FOR BUSINESS DEVELOPMENT**  
**PENGUNAAN MODEL MACHINE LEARNING UNTUK**  
**MEMREDIKSI *CHURN* NASABAH KARTU KREDIT PERBANKAN**  
**(STUDI KASUS DI PT COURSE-NET BANGUN INDONESIA)**

**Oleh :**

**Faisal Akbar Kusprianto / 302210009**

Disetujui dan disahkan sebagai  
**LAPORAN KERJA PRAKTIK**

Bandung, Januari 2025  
Direktur PT Course-Net Indonesia.



**Fransiskus Alvin Winata**

## ABSTRAKSI

*Penelitian ini bertujuan untuk mengembangkan model prediktif guna memprediksi risiko churn nasabah kartu kredit menggunakan pendekatan machine learning. Churn nasabah merupakan tantangan signifikan dalam industri perbankan yang dapat berdampak negatif pada pendapatan dan reputasi bank. Penelitian dilakukan dengan menganalisis dataset yang mencakup informasi demografis dan perilaku penggunaan kartu kredit dari sejumlah nasabah. Metodologi penelitian meliputi beberapa tahapan kritis, yaitu Exploratory Data Analysis (EDA), preprocessing data, dan pengembangan model prediktif. Dua algoritma machine learning utama digunakan: Logistic Regression dan Random Forest. Proses analisis dimulai dengan eksplorasi data untuk mengidentifikasi pola dan hubungan antarvariabel, dilanjutkan dengan preprocessing data yang meliputi one-hot encoding untuk variabel kategorikal dan penanganan data. Hasil penelitian menunjukkan bahwa model Random Forest memberikan performa terbaik dengan nilai ROC AUC 0,88 dan F1 Score 0,81. Analisis mengungkapkan bahwa rasio pemakaian kredit tinggi dan periode tidak aktif adalah faktor kunci yang berkontribusi terhadap risiko churn nasabah. Berdasarkan temuan tersebut, penelitian merekomendasikan sejumlah strategi untuk bank, seperti pemberian layanan prioritas, kampanye re-engagement, dan penawaran solusi pengelolaan kredit yang lebih fleksibel. Penelitian ini memberikan kontribusi penting dalam mengaplikasikan teknologi machine learning untuk mendukung pengambilan keputusan strategis di industri perbankan, khususnya dalam upaya menurunkan tingkat churn nasabah.*

**Kata Kunci :** *Churn Nasabah, Machine Learning, Random Forest, Kartu Kredit, Prediksi Risiko, Analisis Data.*

## KATA PENGANTAR

Dengan memanjatkan puji syukur kehadirat Allah SWT. Tuhan Yang Maha Pengasih dan Penyayang yang telah melimpahkan segala rahmat, hidayah, serta innayah-Nya atas terselesaikannya Kerja Praktik (KP) di PT Course-Net Bangun Indonesia. Laporan disusun untuk memenuhi persyaratan Laporan Kerja Praktik di PT Course-Net Bangun Indonesia dan Fakultas Teknologi Informasi, Program Studi Sistem Informasi, Universitas Bale Bandung. Tujuan dari laporan kerja praktik adalah untuk melaporkan semua kegiatan yang telah dilakukan selama dilaksanakannya program Kerja Praktik (KP) di PT Course-Net Bangun Indonesia. Dalam penyusunan laporan ini, tentu tidak lepas dari pengarahan dan bimbingan dari berbagai pihak. Oleh karena itu, saya ingin mengungkapkan rasa hormat dan terima kasih kepada semua pihak yang telah membantu. Berikut adalah beberapa pihak yang terlibat dalam dilaksanakannya kegiatan Kerja Praktik (KP) ini :

1. Dr. Ir. H. M. Ibrahim Danuwikarsa, M.S. selaku Rektor Universitas Bale Bandung.
2. Yudi Herdiana, S.T., M.T. selaku Dekan Fakultas Teknologi Informasi.
3. Rosmalina, S.T., M.Kom. selaku Ketua Prodi Sistem Informasi dan Dosen Pembimbing Kerja Praktik.
4. Fransiskus Alvin Winata, selaku Direktur PT Course-Net Indonesia.
5. Reynold Matua Sinambela, selaku Mentor Kelas Program Data Science For Business Development.
6. Teman – Teman Kelas Program Data Science For Business Development dan semua pihak yang telah membantu penulis dalam menyelesaikan Program Studi Independen Bersertifikat ini yang tidak dapat penulis sebutkan satu persatu.
7. Keluarga yang selalu memberi semangat dan dukungan.

Berkat kerjasama yang baik dari semua pihak yang telah disebutkan sebelumnya, penulis dapat menyelesaikan laporan kerja praktik ini dengan sebaik mungkin. Meskipun laporan ini belum sempurna, kritik dan saran dari para pembaca sangat diharapkan dan akan sangat bermanfaat untuk kemajuan penulis di masa depan. Sekali lagi, penulis mengucapkan terima kasih yang sebesar-besarnya. Semoga laporan ini memberikan manfaat bagi kita semua.

Bandung, 23 Januari 2025

**Faisal Akbar Kusprianto**

302210009

## DAFTAR ISI

ABSTRAKSI .....	iv
KATA PENGANTAR.....	v
DAFTAR ISI .....	vii
DAFTAR TABEL.....	viii
DAFTAR GAMBAR .....	ix
BAB I PENDAHULUAN .....	1
I.1 Latar Belakang .....	1
I.2 Lingkup .....	3
I.3 Rumusan Masalah .....	4
I.4 Tujuan.....	4
BAB II LINGKUNGAN KERJA PRAKTIK .....	5
II.1 Struktur Organisasi .....	5
II.2 Lingkup Pekerjaan .....	6
II.3 Deskripsi Pekerjaan .....	8
II.4 Jadwal Kerja .....	8
BAB III TEORI PENUNJANG KERJA PRAKTIK.....	13
III.1 Teori Penunjang.....	13
III.2 Peralatan Pengembangan Aplikasi .....	17
BAB IV PELAKSANAAN KERJA PRAKTIK .....	24
IV.1 Input.....	24
IV.2 Proses.....	31
IV.3 Pencapaian Hasil.....	59
BAB V PENUTUP.....	61
V.1 Kesimpulan Dan Saran Mengenai Pelaksanaan.....	61
V.1.1 Kesimpulan Pelaksanaan Kerja Praktik .....	61
V.1.2 Pelaksanaan Kerja Praktik .....	62
V.2 Kesimpulan Dan Saran Mengenai Substansi .....	63
V.2.1 Kesimpulan .....	63
V.2.2 Saran .....	63
DAFTAR PUSTAKA.....	65
LAMPIRAN A .....	68
LAMPIRAN B .....	70
LAMPIRAN C .....	80



## **DAFTAR TABEL**

Table 1 Struktur Organisasi CourseNet Indonesia .....	5
Table 2 Jadwal Kelas Data Science for Business Development .....	8

## DAFTAR GAMBAR

Gambar 1 Credit Card Churn - Dataset.....	26
Gambar 2 Kode untuk menampilkan Informasi Demografis .....	26
Gambar 3 Output Kode untuk menampilkan Informasi Demografis.....	27
Gambar 4 Kode untuk menampilkan Status Churn (Label).....	29
Gambar 5 Output Kode untuk menampilkan Status Churn (Label).....	29
Gambar 6 Import Library .....	31
Gambar 7 Kode untuk membaca & menampilkan data CSV .....	32
Gambar 8 Output Kode untuk membaca & menampilkan data CSV .....	33
Gambar 9 Kode Visualisasi Data Kategorikal .....	34
Gambar 10 Output Kode Visualisasi Data Kategorikal .....	36
Gambar 11 Kode gabungan Data One-Hot Encoded & Non-One Hot Encoded .....	36
Gambar 12 Output gabungan Data One-Hot Encoded & Non-One-Hot Encoded.....	38
Gambar 13 Kode Visualisasi Distribusi Variabel Kategorikal .....	39
Gambar 14 Output Visualisasi Distribusi Variabel Kategorikal.....	40
Gambar 15 kode one-hot encoding menggunakan Python & library scikit-learn .....	41
Gambar 16 Output one-hot encoding menggunakan Python & library scikit-learn.....	44
Gambar 17 Kode untuk memahami Korelasi Data .....	44
Gambar 18 Correlation Matrix Heatmap .....	46
Gambar 19 Mencari & Visualisasi 5 Fitur Terkorelasi dengan Label.....	48
Gambar 20 Top 5 Feature Correlations with label .....	50
Gambar 21 Kode Memprediksi "Label" dengan Logistic Regression & Random Forest .....	52
Gambar 22 Output Hasil Prediksi "Label" dengan Logistic Regression & Random Forest.....	55
Gambar 23 Kode Memprediksi Churn Kartu Kredit dengan Machine Learning .....	56
Gambar 24 Output hasil memprediksi Churn Kartu Kredit dengan Machine Learning .....	58
Gambar 25 Komponen Penilaian Kelas SIB Data Science.....	80
Gambar 26 Kegiatan Kelas SIB Data Science .....	80
Gambar 21 Sesi Mentoring dan Konsultasi .....	81
Gambar 28 Presentasi Days Final Project Kelas SIB Data Science .....	81
Gambar 29 Graduation Days SIB Course-Net Indonesia .....	82
Gambar 30 Learning Management System (LMS) ITBOX.....	83
Gambar 31 Grup WhatsApp SIB Data Science 0224 .....	84
Gambar 32 Grup WhatsApp Konsultasi & Mentoring SIB Data Science ...	84

# **BAB I**

## **PENDAHULUAN**

### **I.1 Latar Belakang**

PT. Course Net Bangun Indonesia menyelenggarakan program Studi Independen Bersertifikat (SIB) dengan fokus pada *Data Science for Business Development*. Program ini dirancang untuk memberikan mahasiswa pengalaman praktis dalam menerapkan ilmu data *science* untuk pengembangan bisnis, yang sangat relevan di era digital saat ini. Program Studi Independen Bersertifikat (SIB) adalah program yang dicanangkan oleh Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Indonesia dalam kerangka Merdeka Belajar – Kampus Merdeka (MBKM). Program ini bertujuan untuk memberikan mahasiswa pengalaman praktis di dunia industri, sehingga mereka dapat mengembangkan keterampilan dan pengetahuan yang relevan dengan kebutuhan pasar kerja.

Layanan kartu kredit telah menjadi salah satu pilar penting dalam industri perbankan, menawarkan kemudahan transaksi sekaligus menjadi sumber pendapatan utama. Dengan meningkatnya persaingan antarbank dan perubahan preferensi nasabah, pengelolaan hubungan dengan nasabah menjadi semakin penting. Dalam konteks ini, memahami kebutuhan dan perilaku nasabah menjadi kunci untuk mempertahankan mereka dalam ekosistem layanan perbankan, khususnya layanan kartu kredit. Namun, mempertahankan nasabah bukanlah tugas yang mudah, terutama di tengah perubahan kondisi ekonomi, preferensi digitalisasi, dan semakin banyaknya alternatif layanan keuangan.

Salah satu tantangan terbesar yang dihadapi oleh tim bisnis di sebuah bank adalah meningkatnya tingkat *churn*, yaitu keputusan nasabah untuk menutup layanan kartu kredit. *Churn* tidak hanya berdampak pada penurunan pendapatan bank, tetapi juga mencerminkan adanya celah dalam memahami dan memenuhi kebutuhan nasabah. Jika dibiarkan tanpa solusi, tingkat *churn* yang tinggi dapat memengaruhi stabilitas bisnis dan reputasi bank di mata nasabah lainnya. Oleh karena itu, masalah *churn* ini menjadi perhatian utama bagi bank untuk segera ditangani.

Untuk menghadapi tantangan ini, diperlukan pendekatan strategis berbasis data yang memanfaatkan teknologi modern seperti machine learning. Dengan menggunakan machine learning, bank dapat membangun model prediktif yang mampu mengidentifikasi nasabah dengan risiko *churn* tinggi berdasarkan pola transaksi, perilaku pembayaran, dan data profil mereka. Informasi ini memungkinkan tim bisnis untuk memberikan intervensi yang tepat sasaran, seperti penawaran khusus atau layanan yang disesuaikan, sehingga dapat mendorong nasabah untuk tetap menggunakan layanan kartu kredit mereka. Pendekatan berbasis teknologi ini diharapkan tidak hanya mengurangi churn, tetapi juga meningkatkan efisiensi strategi retensi nasabah.

Penelitian dari (Teng & Lee, 2019) berfokus pada prediksi risiko gagal bayar (default) oleh pemegang kartu kredit dengan menggunakan lima metode machine learning, yaitu K-Nearest Neighbors (KNN), Decision Tree (DT), Boosting, Support Vector Machine (SVM), dan Neural Network (NN). Penelitian ini bertujuan untuk mengevaluasi performa berbagai metode dalam memprediksi default menggunakan dataset dengan 23 fitur dari UC Irvine Machine Learning Repository. Hasilnya menunjukkan bahwa metode decision tree memberikan akurasi terbaik dalam memprediksi default dibandingkan metode lainnya. Di sisi lain, penelitian dari (Wang et al., 2019) berfokus pada prediksi churn pelanggan di industri dana internet menggunakan metode inovatif Feature Embedded Convolutional Neural Network (FE-CNN). Model ini mengintegrasikan data perilaku dinamis pelanggan dan data demografi statis untuk menghasilkan prediksi yang lebih akurat dibandingkan metode tradisional seperti Logistic Regression, SVM, Random Forest, dan NN. Kedua penelitian memiliki kesamaan dalam memanfaatkan teknologi machine learning untuk memprediksi perilaku spesifik berdasarkan dataset yang besar dan kompleks. Selain itu, keduanya bertujuan meningkatkan akurasi prediksi dengan membandingkan metode yang digunakan. Namun, penelitian (Teng & Lee, 2019) lebih berfokus pada risiko gagal bayar di sektor perbankan, sementara (Wang et al., 2019) menargetkan prediksi churn pelanggan di industri dana internet dengan pendekatan berbasis deep learning. Perbedaan lain terletak pada jenis data yang digunakan; penelitian (Teng & Lee, 2019) menggunakan atribut statis seperti profil pelanggan dan histori pembayaran,

sedangkan (Wang et al., 2019) menggabungkan data dinamis dan statis untuk menghasilkan prediksi yang lebih akurat. Penelitian tentang prediksi churn nasabah kartu kredit perbankan memiliki kesamaan signifikan dengan pendekatan (Wang et al., 2019) yang berfokus pada prediksi churn. Penelitian ini juga memprioritaskan eksplorasi data (exploratory data analysis) dan evaluasi berbagai algoritma machine learning untuk menghasilkan model prediksi terbaik, yang mencerminkan metodologi pada kedua penelitian ini.

## **I.2 Lingkup**

Penelitian ini akan mencakup ruang lingkup berikut untuk mencapai tujuan yang diinginkan :

1. Identifikasi Masalah dan Tujuan Proyek

Menentukan faktor-faktor utama yang berkontribusi terhadap churn nasabah kartu kredit berdasarkan analisis data dari dataset yang tersedia. Hal ini bertujuan untuk memahami pola churn dan menetapkan tujuan pembuatan model prediktif yang efektif.

2. Exploratory Data Analysis (EDA)

Melakukan eksplorasi data mendalam untuk mengidentifikasi pola, hubungan antarvariabel, dan insight penting dari dataset. Proses ini mencakup visualisasi data guna membantu memahami distribusi data dan tren perilaku nasabah.

3. Data Preprocessing

Mempersiapkan data sebelum membangun model, termasuk penanganan nilai kosong (missing values), encoding variabel kategori, normalisasi data, serta pembagian dataset menjadi data latih (training) dan uji (testing).

4. Pengembangan Model Machine Learning

Membuat dan menguji beberapa algoritma machine learning untuk memprediksi churn nasabah. Model yang dikembangkan akan dievaluasi berdasarkan performanya menggunakan metrik seperti akurasi, precision, recall, dan F1-score.

#### 5. Rekomendasi Berdasarkan Model Terbaik

Memberikan rekomendasi strategis berdasarkan hasil prediksi dari model dengan performa terbaik, sehingga bank dapat merancang intervensi yang efektif untuk mengurangi churn nasabah kartu kredit.

### **I.3 Rumusan Masalah**

Berdasarkan latar belakang yang telah dikemukakan di atas, maka yang menjadi pokok masalah dalam penelitian ini adalah sebagai berikut :

1. Apa saja faktor utama yang memengaruhi churn nasabah kartu kredit di dataset yang digunakan?
2. Bagaimana hasil exploratory data analysis (EDA) dapat memberikan wawasan tentang perilaku dan pola churn nasabah?
3. Apa langkah-langkah preprocessing yang perlu dilakukan untuk mempersiapkan data sebelum pengembangan model?
4. Algoritma machine learning mana yang memberikan performa terbaik dalam memprediksi churn nasabah?
5. Bagaimana rekomendasi yang dapat diimplementasikan berdasarkan hasil prediksi model terbaik?

### **I.4 Tujuan**

Berdasarkan rumusan masalah di atas, maka tujuan dalam penelitian ini adalah untuk :

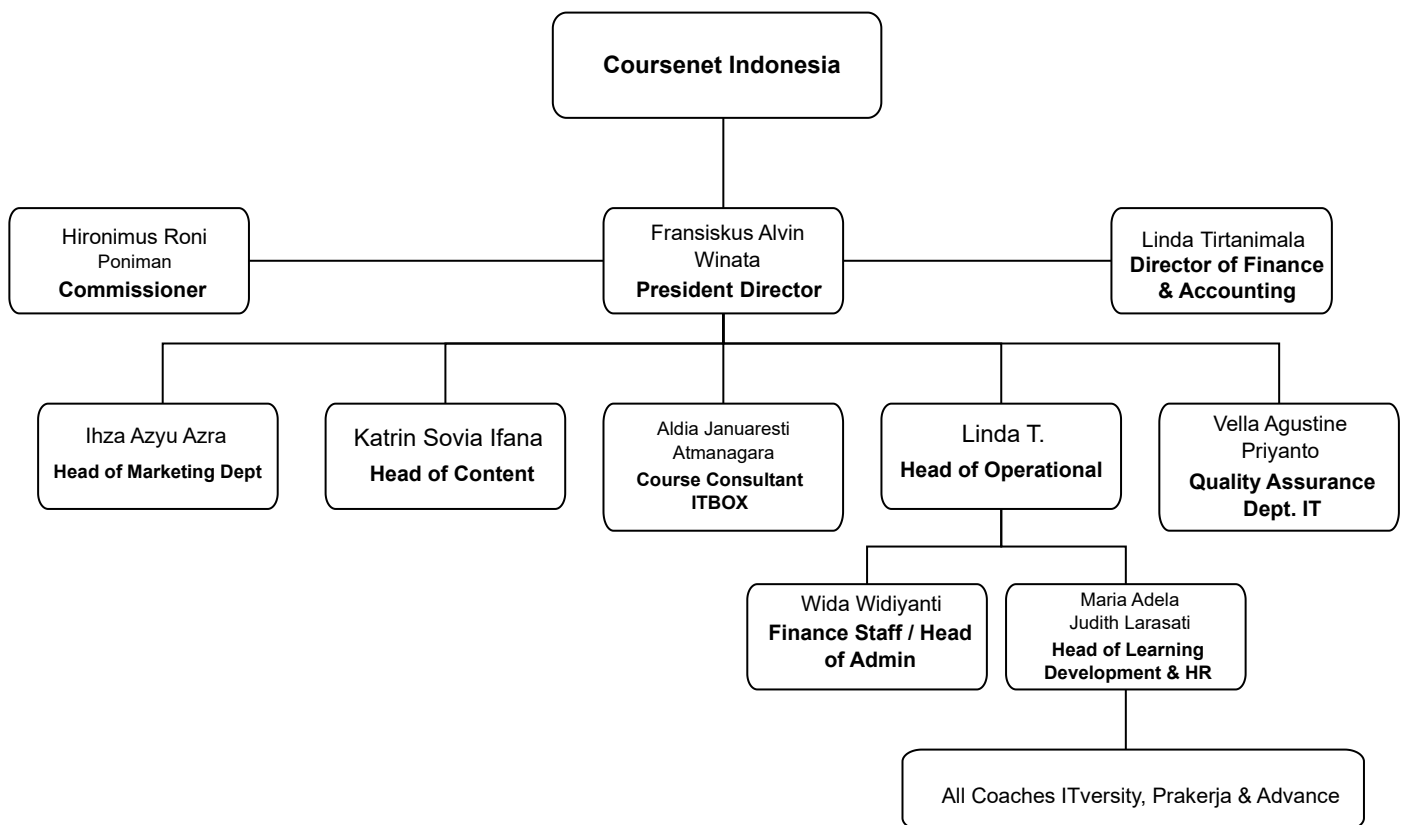
1. Mengidentifikasi faktor-faktor utama yang berpengaruh terhadap churn nasabah kartu kredit melalui analisis dataset.
2. Melakukan exploratory data analysis (EDA) untuk menemukan insight dan pola data yang relevan dengan churn nasabah.
3. Melaksanakan proses data preprocessing untuk memastikan data siap digunakan dalam pengembangan model.
4. Mengembangkan dan mengevaluasi model machine learning yang efektif untuk memprediksi churn nasabah kartu kredit.
5. Memberikan rekomendasi berbasis hasil prediksi untuk mendukung pengambilan keputusan dalam mengurangi churn nasabah kartu kredit di perbankan.

## BAB II

### LINGKUNGAN KERJA PRAKTIK

#### II.1 Struktur Organisasi

*Table 1 Struktur Organisasi CourseNet Indonesia*



1. **President Director**  
Memimpin dan mengelola perusahaan secara keseluruhan, bertanggung jawab atas pengambilan keputusan strategis, serta menjadi komunikator utama antara perusahaan dan pemangku kepentingan.
2. **Commisioner**  
Melakukan pengawasan terhadap kinerja direksi dan memberikan nasihat strategis.
3. **Director of Finance & Accounting**  
Mengelola aspek keuangan dan akuntansi perusahaan, termasuk perencanaan keuangan, penganggaran, dan pelaporan keuangan.

4. Head of Marketing Dept  
Mengembangkan strategi pemasaran untuk meningkatkan brand awareness dan penjualan produk atau layanan.
5. Head of Content  
Bertanggung jawab atas pengembangan konten yang relevan dan menarik untuk audiens target.
6. Course Consultant ITBOX  
Memberikan konsultasi terkait kursus teknologi informasi kepada klien.
7. Head of Operational  
Mengelola operasi sehari-hari perusahaan untuk memastikan efisiensi dan efektivitas.
  - a. Finance Staff / Head of Admin  
Menangani administrasi keuangan dan mendukung kegiatan operasional.
  - b. Head of Learning Development & HR  
Mengembangkan program pelatihan dan pengembangan untuk karyawan.
    - a) All Coaches ITversity, Prakerja & Advance  
Mengajar dan melatih peserta kursus dalam bidang teknologi informasi.
8. Quality Assurance Dept IT  
Memastikan kualitas produk atau layanan yang dihasilkan oleh perusahaan sesuai dengan standar yang ditetapkan.

## **II.2 Lingkup Pekerjaan**

Dalam program Kerja Praktik di PT. Course Net Bangun Indonesia, saya adalah seorang mentee yang mengambil kelas *Data Science for Business Development*. Program ini berlangsung dari tanggal 16 Februari hingga 30 Juni 2024 selama 5 bulan. Saya dibimbing oleh seorang fasilitator yang berperan sebagai coach, yang menjelaskan materi dan memberikan panduan selama proses pembelajaran. Tanggung jawab saya dalam pekerjaan ini adalah mempelajari dan menyelesaikan tugas-tugas yang diberikan oleh Course Net. Selain itu, saya juga mengerjakan proyek akhir yang diberikan oleh mitra dengan tepat waktu.



1. Sesi Belajar Sistem Online (WEBINAR)

Sesi ini merupakan sesi di mana peserta didik belajar dan mendapatkan pemahaman teori dan praktik lengkap tentang *Data Science for Business Development* sesuai dengan kurikulum yang telah ditentukan. Kegiatan pembelajaran ini melibatkan coach dan semua peserta dalam satu kelas. Pengajaran ini dilakukan secara daring melalui Platform *Zoom Meeting*. Sesi ini terdapat juga di mana coach akan memberikan sejumlah tugas latihan kepada seluruh mahasiswa setelah selesai dengan materi. Hal ini dilakukan sebagai bagian dari ulasan materi yang telah diajarkan, sehingga mahasiswa dapat mempertahankan pemahaman terhadap materi-materi sebelumnya.

2. *Learning Management System (LMS)*

Sesi ini merupakan sesi di mana peserta didik belajar dan mendapatkan pemahaman teori secara online melalui platform ITBOX yang terdapat Sembilan topik utama materi *Data Science for Business Development* sesuai dengan kurikulum yang telah ditentukan. Peserta didik disini belajar teori melalui video learning dengan diakhiri mengerjakan kuis dan setelah menyelesaikan pertopik utama peserta didik harus menyelesaikan ujian akhir untuk mendapatkan sertifikat penyelesaian.

3. Sesi Mentoring

Sesi ini merupakan dimana para peserta didik mendapatkan sesi mentoring berkaitan dengan topik pembelajaran dari Sesi Belajar Sistem Online (WEBINAR) maupun dari sesi *Learning Management System (LMS)*. Skema mentoring yang telah ditetapkan oleh course net dimulai dari peserta didik Dibuatkan grup *WhatsApp* 25 orang peserta. Sesi ini dilakukan secara daring melalui Platform *Zoom Meeting* dan *WhatsApp*.

4. Sesi Konsultasi

Sesi ini sama halnya dengan sesi mentoring pada sesi ini para peserta didik mendapatkan sesi konsultasi. Sesi ini dilaksanakan Sebulan sekali untuk bertemu dengan Mentor dan Dosen Pendamping Program (DPP) dengan durasi 1 sampai 3 jam. Peserta didik tidak wajib hadir. Sesi ini bisa digunakan untuk bertanya secara langsung, topik mengenai pelaksanaan

kampus Merdeka maupun diluar kampus Merdeka. Sesi ini dilakukan secara daring melalui Platform *Zoom Meeting* dan WhatsApp.

#### 5. Final Project

Sesi ini didedikasikan khusus untuk mengerjakan proyek akhir. Dalam sesi ini, mentor akan memberikan tugas kepada seluruh mahasiswa di kelas *Data Science for Business Development* untuk menganalisis sebuah studi kasus dengan batas waktu pengerjaan yang telah ditentukan dan diakhir diharuskan mempresentasikan hasil analisis nya secara langsung.

### II.3 Deskripsi Pekerjaan

Pada proyek akhir ini, penulis melakukan analisis masalah mengenai bagaimana memprediksi *churn* nasabah kartu kredit menggunakan model *machine learning* dan mengevaluasi performanya. Penulis diberikan file dataset *Credit Card Churn - Objective & Description* oleh Course Net Indonesia. Dari dataset ini, Penulis diminta untuk mencari pola customer yang *churn* dan yang tidak supaya perusahaan dapat mencegah bertambahnya jumlah nasabah yang *churn*.

### II.4 Jadwal Kerja

Penulis telah mengikuti pembelajaran dari kelas *Data Science for Business Development* selama 5 bulan (32 sesi, 3 jam setiap sesi) dari Coach profesional yang ahli di bidangnya. Pembelajaran ini menggunakan metode *live lecture* dan *self-study learning*. *Live lecture* dilakukan secara sinkronus dalam tiga pertemuan hari Senin, Rabu, dan Jum'at dimulai pukul 19:00 hingga 22:00 WIB. *self-study learning* berupa video online menggunakan platform resmi Course-Net Indonesia yaitu ITBOX. Berikut merupakan jadwal dari kelas *Data Science for Business Development* secara rinci :

Table 2 Jadwal Kelas Data Science for Business Development

Month	Main Topic	Date	Time	Sub Topic	Online Learning	Video Online
Month 1 (Feb s/d Mar)	ALGO	Senin, 19 Februari 2024	19:00 sd 22:00 WIB	I/O Syntax and variable, Arithmetic	Week 1 - IT BOX ALGO	- Pengantar Program Algoritma Sintaks Input

				Operation, Selection	(Self Study)	Output Variable Operasi Aritmatika
		Rabu, 21 Februari 2024	19:00 sd 22:00 WIB	Repetition, Array & Pointer, Function		- Selection Repetition - Looping Lanjutan - Bangun Datar Array dan Pointer
		Jumat, 23 Februari 2024	19:00 sd 22:00 WIB	Built-in function, Struct		- Function dan Built-in function Recursive and Struct - Sorting dan Searching Latihan Soal Menu + Searching Bahas Soal Coding Interview
	OOP	Senin, 26 Februari 2024	19:00 sd 22:00 WIB	Intro java wrapper class and method	Week 2 - IT BOX OOP (Self Study)	- Pengantar Program Object-Oriented Programming Introduction to Java
		Rabu, 28 Februari 2024	19:00 sd 22:00 WIB	Array arraylist vector oop concept		- Wrapper Class dan Method
		Jumat, 1 Maret 2024	19:00 sd 22:00 WIB	Inheritance		- Array, ArrayList, Vector - OOP Concept Inheritance - Polymorphism, Kesimpulan
	Database	Senin, 4 Maret 2024	19:00 sd 22:00 WIB	DDL DML	Week 3 - IT BOX DB (Self Study)	- Pendahuluan Membuat dan Memodifikasi Table
		Rabu, 6 Maret 2024	19:00 sd 22:00 WIB	Grouping Aggregate Order By String and Date Function, Join Union		- Select Statement Dasar

		Jumat, 8 Maret 2024	19:00 sd 22:00 WIB	Normalization		
	Basic Network	Rabu, 13 Maret 2024	19:00 sd 22:00 WIB	Intro computer network & pengenalan IP	Week 4 - IT BOX BN (Self Study)	<ul style="list-style-type: none"> <li>- Pengantar Program Basic Network &amp; Introduction to Computer Network</li> <li>- Pengenalan IP Basic Advanced Subnetting</li> <li>- Static Routing VLAN (Virtual Local Area Network)</li> <li>- Dynamic Routing Access List</li> <li>- WLAN (Wireless Local Area Network) &amp; Kesimpulan</li> </ul>
		Kamis, 14 Maret 2024	19:00 sd 22:00 WIB	Basic advanced subnetting, Static routing, VLAN		
		Jumat, 15 Maret 2024	19:00 sd 22:00 WIB	Dynamic Routing ACL		
Month 2 (Maret sd April)	Data Science	Senin, 18 Maret 2024	19:00 sd 22:00 WIB	Introduction to Data Science	Week 1 - IT BOX DB (Self Study)	<ul style="list-style-type: none"> <li>- Perhitungan di SQL Joins</li> <li>- Union Subqueryist</li> </ul>
		Rabu, 20 Maret 2024	19:00 sd 22:00 WIB	R Programming and visualization in Data Science		
		Jumat, 22 Maret 2024	19:00 sd 22:00 WIB	Lab R Programming and visualization in Data Science		
		Senin, 25 Maret 2024	19:00 sd 22:00 WIB	Python	Week 2 - IT BOX DB (Self Study)	<ul style="list-style-type: none"> <li>- Case When Over Partition Statement</li> <li>- View Store Procedure</li> </ul>
		Sabtu, 27 April 2024	09:00 sd 16:00 WIB	Lab Python		
		Senin, 1 April 2024	19:00 sd 22:00 WIB	Introduction to Regression	Week 3 - IT BOX	<ul style="list-style-type: none"> <li>- Trigger</li> </ul>

		Rabu, 3 April 2024	19:00 sd 22:00 WIB	Lab Introduction to Regression	DB (Self Study)	- Membaca & Memahami SQL Query yang sudah ada
		Jumat, 5 April 2024	19:00 sd 22:00 WIB	Introduction to Classification		
Month 3 (April sd Mei)	Data Science	Rabu, 17 April 2024	19:00 sd 22:00 WIB	Lab Introduction to Classification	Week 1 - IT BOX	- Perkenalan Data Science & Berkenalan dengan r
		Jumat, 19 April 2024	19:00 sd 22:00 WIB	Introduction to Hierarchical Clustering	DS (Self Study)	- Perkenalan Machine Learning & Machine Learning Lainnya
		Senin, 22 April 2024	19:00 sd 22:00 WIB	Introduction to Association Rules	Week 2 - IT BOX DS (Self Study)	- Perkenalan Machine Learning & Machine Learning Lainnya
		Rabu, 24 April 2024	19:00 sd 22:00 WIB	Introduction to Data Preparation		- Pengumpulan Data, Pembangunan Fitur & Pembersihan Data
		Senin, 29 April 2024	19:00 sd 22:00 WIB	Introduction to Deep Learning and Tensorflow	Week 3 - IT BOX DS (Self Study)	- Pemilihan Fitur & Pengembangan Model, Mengelola Data yang Tidak Seimbang
		Jumat, 3 Mei 2024	19:00 sd 22:00 WIB	Computer Vision		- Pengenalan Python
		Senin, 6 Mei 2024	19:00 sd 22:00 WIB	Lab Computer Vision	Week 4 - IT BOX DS (Self Study)	- Natural Language Processing
		Rabu, 8 Mei 2024	19:00 sd 22:00 WIB	NLP		- Deep Learning
Month 4 (Mei sd Juni)	Data Science	Senin, 13 Mei 2024	19:00 sd 22:00 WIB	Lab NLP	Week 1 - IT BOX	- Bunga Rampai Data Science

		Rabu, 15 Mei 2024	19:00 sd 22:00 WIB	Time Series	DS (Self Study)	
		Kamis, 16 Mei 2024	19:00 sd 22:00 WIB	Exam ITBOX		
		Jumat, 17 Mei 2024	19:00 sd 22:00 WIB	Lab Time Series		
		20 Mei - 07 Juni 2024	Project Data Science			
Month 5 (Juni)	Data Science	10 Juni - 14 Juni 2024				
		19 Juni - 28 Juni 2024	Presentasion Day			

## **BAB III**

### **TEORI PENUNJANG KERJA PRAKTIK**

#### **III.1 Teori Penunjang**

##### **1. Churn Nasabah Pada Industri Perbankan**

Churn nasabah di industri perbankan dipengaruhi oleh berbagai faktor, termasuk kualitas layanan pelanggan, kepuasan, kepercayaan, dan biaya pengalihan, yang sangat penting dalam mempertahankan klien di pasar yang kompetitif (Geeta Rani & Dr. Asha ., 2024) (Dangiso & Dangiso Dakucho, 2024). Teori perilaku konsumen menjelaskan perilaku churn melalui lensa loyalitas pelanggan, menekankan bahwa pengalaman positif dan hubungan dengan bank menumbuhkan loyalitas, sementara pengalaman negatif menyebabkan pengurangan (Ali; Noraei & Kavosh, 2021).

Selain itu, faktor demografis seperti usia, pendapatan, dan jumlah produk yang dimiliki oleh pelanggan berdampak signifikan pada tingkat churn, dengan segmen pelanggan tertentu menunjukkan kecenderungan keberangkatan yang lebih tinggi (Wen, 2023) (Oetama, 2023). Interaksi faktor-faktor ini menyoroti perlunya bank untuk meningkatkan penyampaian layanan dan strategi keterlibatan pelanggan untuk mengurangi churn dan mempromosikan loyalitas jangka Panjang (Dangiso & Dangiso Dakucho, 2024) (Ali; Noraei & Kavosh, 2021).

##### **2. Machine Learning**

###### **a. Teori Dasar Machine Learning**

Machine learning adalah cabang dari kecerdasan buatan yang berfokus pada pengembangan algoritma dan model yang memungkinkan komputer untuk belajar dari dan membuat prediksi atau keputusan berdasarkan data.

Machine learning dapat dibagi menjadi beberapa kategori, termasuk :

- 1) Supervised Learning : Model dilatih menggunakan data yang sudah dilabeli.
- 2) Unsupervised Learning : Model dilatih dengan data yang tidak dilabeli, dan tujuannya adalah untuk menemukan pola atau struktur dalam data.

- 3) Reinforcement Learning : Di mana agen belajar melalui interaksi dengan lingkungan untuk mencapai tujuan tertentu.

Langkah - Langkah Alur Proyek Machine Learning sebagai berikut :

- 1) Mengumpulkan data
- 2) Menyiapkan data (pembersihan dan pemrosesan)
- 3) Melatih model
- 4) Menguji dan mengevaluasi model
- 5) Mengimplementasikan dan memelihara model.

Dalam machine learning, tantangan utama termasuk jumlah data yang tidak mencukupi, data yang tidak representatif, dan masalah overfitting atau underfitting (Géron, 2019).

#### **b. Model Machine Learning**

Teori dan prinsip dasar algoritma Random Forest merupakan metode ensemble yang menggabungkan beberapa pohon keputusan (decision trees) untuk meningkatkan akurasi klasifikasi. Algoritma ini bekerja dengan cara membangun sejumlah pohon keputusan, di mana setiap pohon dibentuk berdasarkan subset acak dari data pelatihan dan fitur. Proses ini melibatkan dua langkah utama : pemilihan acak data (bagging) dan pemilihan fitur acak pada setiap node saat membangun pohon. Dengan cara ini, Random Forest mengurangi risiko overfitting yang sering terjadi pada model tunggal, karena setiap pohon dapat memberikan suara untuk kelas yang paling umum berdasarkan hasil klasifikasinya. (Jin et al., 2020).

Salah satu aspek penting dari Random Forest adalah konsep margin, yang mengukur seberapa baik model dapat memisahkan kelas-kelas yang berbeda. Margin yang lebih besar menunjukkan kepercayaan yang lebih tinggi dalam klasifikasi. Algoritma ini juga menggunakan estimasi out-of-bag untuk memantau kesalahan, kekuatan, dan korelasi antar pohon. Dengan cara ini, Random Forest dapat memberikan estimasi kesalahan yang akurat tanpa memerlukan set data terpisah untuk pengujian. Keunggulan Random Forest dibandingkan dengan metode lain seperti Adaboost adalah kemampuannya untuk menangani variabel input yang sangat banyak dan sering kali memiliki informasi yang terbatas. Dalam



banyak aplikasi, seperti diagnosis medis atau pengenalan karakter, Random Forest menunjukkan akurasi yang lebih baik dengan menggabungkan hasil dari banyak pohon yang dibangun dengan fitur acak. Hal ini menjadikan Random Forest sebagai salah satu algoritma yang sangat efektif dalam berbagai tugas pembelajaran mesin, baik untuk klasifikasi maupun regresi (Jin et al., 2020).

Regresi logistik adalah metode statistik yang digunakan untuk klasifikasi biner, memprediksi probabilitas hasil biner (0 atau 1) berdasarkan satu atau lebih variabel prediktor. Variabel dependen dalam regresi logistik adalah biner, artinya dapat menghasilkan dua kemungkinan hasil. Ini sangat penting untuk aplikasi seperti prediksi churn pelanggan, di mana hasilnya adalah apakah pelanggan akan pergi atau tinggal. Regresi logistik memodelkan hubungan antara variabel dependen dan variabel independen menggunakan fungsi logistik. Fungsi ini mengubah kombinasi linear input menjadi probabilitas yang berkisar antara 0 dan 1, sehingga cocok untuk hasil biner. Model mengasumsikan bahwa variabel independen terkait secara linier dengan peluang log dari variabel dependen. Itu tidak memerlukan variabel dependen untuk didistribusikan secara normal, yang merupakan perbedaan utama dari regresi linier. Efektivitas regresi logistik dapat dievaluasi menggunakan berbagai metrik, termasuk kurva ROC dan uji Hosmer-Lemeshow, yang menilai akurasi prediksi model. Regresi logistik banyak digunakan dalam manajemen risiko, terutama dalam memprediksi churn pelanggan di industri seperti telekomunikasi, di mana memahami perilaku pelanggan sangat penting untuk strategi retensi (Velu, 2021).

Regresi logistik digunakan untuk memprediksi kemungkinan churn pelanggan dengan menganalisis berbagai faktor yang mempengaruhi perilaku pelanggan. Ini membantu dalam mengidentifikasi pelanggan mana yang berisiko meninggalkan perusahaan. Dalam konteks churn pelanggan, variabel dependen adalah biner: ini menunjukkan apakah pelanggan telah churned (1) atau tidak (0). Sifat biner ini menjadikan regresi logistik pilihan yang tepat untuk memodelkan prediksi churn. Model ini menggunakan berbagai variabel independen, seperti demografi pelanggan, pola penggunaan layanan, dan interaksi layanan pelanggan, untuk memprediksi churn. Misalnya, faktor-faktor seperti kualitas panggilan dan rencana layanan telah diidentifikasi sebagai prediktor signifikan churn dalam

penelitian. Efektivitas model regresi logistik dalam memprediksi churn dapat dinilai menggunakan metrik seperti akurasi, presisi, dan ingatan. Metrik ini membantu menentukan seberapa baik model mengidentifikasi pelanggan yang cenderung berpindah dan menginformasikan strategi retensi. Dengan memahami prediktor churn melalui regresi logistik, perusahaan dapat mengembangkan strategi retensi yang ditargetkan, seperti penawaran yang dipersonalisasi atau peningkatan pengiriman layanan, untuk mengurangi tingkat churn dan meningkatkan loyalitas pelanggan (Velu, 2021).

### **3. Pengolahan Data dan Analisis Churn**

Analisis churn merupakan proses penting dalam memahami perilaku pelanggan dan mempertahankan mereka dalam bisnis. Dalam konteks ini, konsep data mining dan pengolahan data berperan krusial. Data mining adalah proses menemukan pola dan informasi berguna dari kumpulan data besar, yang dapat membantu dalam mengidentifikasi pelanggan yang berpotensi untuk churn (Ha et al., 2011). Pengolahan data mencakup langkah-langkah seperti pengumpulan, pembersihan, dan transformasi data, yang semuanya penting untuk memastikan bahwa data yang digunakan dalam analisis churn adalah akurat dan relevan. Selanjutnya, teori feature engineering menjadi penting dalam menciptakan fitur-fitur baru yang dapat meningkatkan kinerja model prediksi churn. Proses ini melibatkan pemilihan, transformasi, dan penciptaan fitur dari data mentah yang dapat memberikan wawasan lebih dalam tentang perilaku pelanggan (Franklin, 2005). Data preprocessing juga tidak kalah penting, karena langkah ini memastikan bahwa data yang digunakan dalam model bebas dari noise dan siap untuk analisis lebih lanjut. Setelah model dibangun, evaluasi model menjadi langkah krusial untuk menilai kinerjanya. Metode evaluasi seperti akurasi, precision, recall, F1-score, dan AUC-ROC memberikan gambaran yang komprehensif tentang seberapa baik model dalam memprediksi churn dengan menggunakan metrik ini, perusahaan dapat mengoptimalkan strategi retensi pelanggan dan mengurangi tingkat churn secara efektif (Saito & Rehmsmeier, 2015).

#### **4. Manajemen Risiko Di Perbankan**

Customer churn secara signifikan berdampak pada profitabilitas bank dengan meningkatkan biaya operasional dan mengurangi aliran pendapatan, karena mempertahankan pelanggan yang sudah ada umumnya lebih murah daripada memperoleh yang baru. Tingkat churn yang tinggi dapat menyebabkan penurunan loyalitas dan kepercayaan pelanggan, yang pada akhirnya mempengaruhi posisi pasar bank dan stabilitas keuangan (Qin, 2024) (POPESCU, 2018). Untuk mengurangi risiko churn, bank dapat mengadopsi strategi teknologi seperti analisis data dan sistem manajemen hubungan pelanggan (CRM), yang memungkinkan layanan yang dipersonalisasi dan keterlibatan proaktif dengan pelanggan (Fomina & Khodkovskaya, 2023) (The et al., 2020). Teknologi ini memfasilitasi identifikasi pelanggan yang berisiko melalui pemodelan prediktif, memungkinkan bank untuk menerapkan strategi retensi yang ditargetkan, seperti penawaran khusus atau peningkatan pengiriman layanan, sehingga meningkatkan kepuasan dan loyalitas pelanggan (POPESCU, 2018). Selain itu, mengintegrasikan kerangka kerja manajemen risiko yang memperhitungkan perilaku pelanggan dapat memperkuat ketahanan bank terhadap kerugian keuangan terkait churn (Qin, 2024) (The et al., 2020).

### **III.2 Peralatan Pengembangan Aplikasi**

#### **1. Perangkat Keras (Hardware)**

Spesifikasi Central Processing Unit (CPU) merupakan komponen utama dalam sistem komputer yang bertanggung jawab menjalankan instruksi dan memproses data. Dalam konteks machine learning, CPU dengan spesifikasi optimal sangat penting untuk mendukung kinerja yang efisien. Pertama, jumlah core pada CPU berperan signifikan, di mana CPU dengan lebih banyak core, minimal 8 core, mampu meningkatkan kinerja dalam pemrosesan data paralel, yang diperlukan untuk algoritma machine learning yang kompleks. Kedua, kecepatan clock CPU juga menjadi faktor penting, di mana kecepatan minimal 3.0 GHz dapat meningkatkan performa terutama untuk tugas-tugas yang tidak dapat diparalelkan. Terakhir, kapasitas cache memory yang besar memungkinkan CPU mengakses data dengan lebih cepat, yang sangat dibutuhkan dalam pengolahan dataset besar.

Dengan spesifikasi ini, CPU dapat mendukung pengolahan machine learning secara optimal (Ian Goodfellow, Yoshua Bengio, 2017).

Spesifikasi Graphics Processing Unit (GPU) memegang peranan penting dalam machine learning, terutama untuk deep learning, karena kemampuannya melakukan komputasi paralel yang efisien. Salah satu faktor utama yang memengaruhi performa GPU adalah jumlah core CUDA, di mana GPU dengan lebih banyak core, minimal 256 core, dapat meningkatkan kecepatan pelatihan model. Selain itu, memori video (VRAM) yang memadai, dengan kapasitas minimal 8 GB, sangat diperlukan untuk menangani model yang besar dan dataset yang kompleks. Faktor lain yang juga penting adalah arsitektur GPU, di mana arsitektur terbaru seperti NVIDIA Ampere atau Turing mampu menawarkan efisiensi dan kinerja yang jauh lebih tinggi dibanding generasi sebelumnya. Dengan spesifikasi ini, GPU dapat mendukung kebutuhan komputasi intensif dalam machine learning secara optimal (Keijsers, 2010).

Spesifikasi Random Access Memory (RAM) memiliki peran penting sebagai memori sementara yang menyimpan data yang sedang diproses. Dalam konteks machine learning, kapasitas RAM yang memadai menjadi salah satu faktor utama untuk memastikan kinerja yang optimal. Kapasitas minimal 16 GB disarankan, tetapi untuk mengolah dataset yang lebih besar, kapasitas 32 GB atau lebih akan lebih ideal. Selain kapasitas, kecepatan RAM juga berpengaruh signifikan terhadap kinerja, di mana RAM dengan kecepatan minimal 3200 MHz dapat meningkatkan throughput data dan mempercepat pemrosesan. Dengan kombinasi kapasitas dan kecepatan yang memadai, RAM dapat mendukung kebutuhan komputasi machine learning secara efisien (Géron, 2019).

Penyimpanan yang cepat dan berkapasitas besar merupakan komponen penting dalam machine learning untuk menyimpan dataset dan model. Tipe penyimpanan yang digunakan sangat berpengaruh pada kinerja, di mana Solid State Drive (SSD) lebih disarankan dibandingkan Hard Disk Drive (HDD) karena memiliki kecepatan baca dan tulis yang lebih tinggi. Hal ini dapat mempercepat waktu pemuatan data, sehingga meningkatkan efisiensi kerja. Selain itu, kapasitas penyimpanan yang memadai juga diperlukan, dengan minimal 1 TB disarankan

untuk mengakomodasi dataset besar dan model yang dilatih. Kombinasi penyimpanan yang cepat dan berkapasitas besar dapat mendukung kebutuhan komputasi machine learning secara optimal (Géron, 2019).

## **2. Perangkat Lunak (Software)**

### **1) Bahasa Pemrograman Python**

Python adalah bahasa pemrograman tingkat tinggi yang dikenal karena sintaksisnya yang sederhana dan kemudahan dalam pembacaan kode. Hal ini menjadikannya pilihan utama bagi para ilmuwan data dan pengembang machine learning. Beberapa alasan mengapa Python sangat populer dalam pengembangan machine learning adalah :

- a. Sintaksis yang Mudah Dipahami : Python memiliki sintaksis yang bersih dan mudah dipahami, sehingga memudahkan pemula untuk belajar dan mengimplementasikan algoritma machine learning (Géron, 2019).
- b. Komunitas yang Besar: Python memiliki komunitas yang aktif, sehingga banyak sumber daya, tutorial, dan forum yang tersedia untuk membantu pengembang (VanderPlas, 2016).
- c. Ekosistem Library yang Kaya: Python mendukung berbagai library yang dirancang khusus untuk pengolahan data dan machine learning, yang memungkinkan pengembang untuk membangun model dengan lebih efisien (McKinney, 2022).

Beberapa library populer yang digunakan dalam pengolahan data dan pelatihan model machine learning:

#### **a. Pandas**

Pandas adalah library yang digunakan untuk manipulasi dan analisis data. Dengan struktur data seperti DataFrame, Pandas memudahkan pengguna untuk melakukan operasi seperti penggabungan, pengelompokan, dan pemfilteran data. Pandas sangat berguna dalam tahap pra-pemrosesan data sebelum model machine learning dilatih (McKinney, 2022).

b. NumPy

NumPy adalah library fundamental untuk komputasi ilmiah di Python. Ini menyediakan dukungan untuk array multidimensi dan berbagai fungsi matematis yang memungkinkan pengolahan data numerik yang efisien. NumPy sering digunakan dalam pengolahan data dan sebagai basis untuk banyak library lain, termasuk Pandas (Oliphant, 2010).

c. Matplotlib

Matplotlib adalah library visualisasi data yang memungkinkan pengguna untuk membuat grafik dan plot yang informatif. Visualisasi data sangat penting dalam machine learning untuk memahami distribusi data, hubungan antar variabel, dan hasil model (Hunter, 2007).

d. Seaborn

Seaborn adalah library visualisasi data yang dibangun di atas Matplotlib. Seaborn menyediakan antarmuka yang lebih sederhana dan lebih estetik untuk membuat visualisasi yang kompleks, seperti heatmaps dan pair plots. Ini sangat berguna untuk eksplorasi data dan analisis statistik (Waskom, 2021).

## 2) Framework dan Library dalam Machine Learning

Framework dan library adalah sekumpulan alat dan fungsi yang dirancang untuk memudahkan pengembang dalam membangun aplikasi machine learning. Mereka menyediakan struktur yang diperlukan untuk mengorganisir kode, serta fungsi-fungsi yang telah dioptimalkan untuk berbagai algoritma machine learning. Penggunaan framework dan library dapat mempercepat proses pengembangan dan mengurangi kemungkinan kesalahan (Adugna et al., 2024).

a. Scikit-learn : Overview

Scikit-learn adalah salah satu library machine learning yang paling populer di Python. Library ini dirancang untuk memberikan alat yang sederhana dan efisien untuk analisis data dan pemodelan. Scikit-learn mendukung berbagai algoritma untuk klasifikasi, regresi, clustering, dan pengurangan dimensi, serta menyediakan alat untuk evaluasi model dan pemilihan fitur (Pedregosa et al., 2011).

### 1) Fitur Utama Scikit-learn

- a) Konsistensi API: Scikit-learn memiliki antarmuka yang konsisten untuk semua algoritma, yang memudahkan pengguna untuk beralih antara model yang berbeda (Raschka, S., & Mirjalili, 2019).
- b) Dokumentasi yang Baik: Scikit-learn dilengkapi dengan dokumentasi yang komprehensif, tutorial, dan contoh yang membantu pengguna baru untuk memahami cara menggunakan library ini (Géron, 2019).
- c) Integrasi dengan Library Lain: Scikit-learn dapat dengan mudah diintegrasikan dengan library lain seperti NumPy, Pandas, dan Matplotlib, yang memungkinkan pengguna untuk melakukan analisis data yang lebih kompleks (Raschka, S., & Mirjalili, 2019).

### 2) Proses Pengembangan Model dengan Scikit-learn

Menurut (Géron, 2019) Proses pengembangan model machine learning dengan Scikit-learn umumnya meliputi beberapa langkah :

- a) Pengumpulan Data : Mengumpulkan data yang relevan untuk analisis.
- b) Pra-pemrosesan Data : Membersihkan dan menyiapkan data untuk analisis, termasuk penanganan nilai yang hilang dan normalisasi.
- c) Pemilihan Model : Memilih algoritma yang sesuai berdasarkan jenis masalah (klasifikasi, regresi, dll.).
- d) Pelatihan Model : Menggunakan data pelatihan untuk melatih model.
- e) Evaluasi Model : Menggunakan metrik evaluasi untuk menilai kinerja model.
- f) Tuning Hyperparameter : Mengoptimalkan parameter model untuk meningkatkan kinerja.
- g) Implementasi : Menggunakan model yang telah dilatih untuk membuat prediksi pada data baru.

### 3) Keunggulan dan Keterbatasan Scikit-learn

Keunggulan Scikit-learn termasuk kemudahan penggunaan, dokumentasi yang baik, dan komunitas yang aktif. Namun, ada juga keterbatasan, seperti kurangnya dukungan untuk deep learning, yang lebih baik ditangani oleh library lain seperti TensorFlow atau PyTorch (Raschka, S., & Mirjalili, 2019).

### **3) Google Colab**

Google Colab adalah platform berbasis cloud yang memungkinkan pengguna untuk menulis dan menjalankan kode Python di browser. Ini sangat berguna untuk analisis data, pembelajaran mesin, dan pengembangan model AI. Teori kolaborasi menekankan pentingnya kerja sama dalam mencapai tujuan bersama. Google Colab memungkinkan beberapa pengguna untuk bekerja pada proyek yang sama secara bersamaan, yang mendukung kolaborasi dalam penelitian dan pengembangan. Menurut (Johnson & Johnson, 2021), kolaborasi dapat meningkatkan pemahaman dan keterampilan individu melalui interaksi sosial dan pertukaran ide. Teori pembelajaran berbasis proyek (Project-Based Learning, PBL) menekankan pembelajaran aktif di mana siswa terlibat dalam proyek nyata. Google Colab menyediakan lingkungan yang ideal untuk PBL, di mana siswa dapat menerapkan teori yang dipelajari dalam praktik. (Zhang et al., 2021) menyatakan bahwa PBL dapat meningkatkan motivasi dan keterlibatan siswa, serta membantu mereka mengembangkan keterampilan kritis dan kreatif. Teori komputasi awan (cloud computing) menjelaskan bagaimana sumber daya komputasi dapat diakses melalui internet. Google Colab adalah contoh dari komputasi awan yang memungkinkan pengguna untuk menjalankan kode tanpa memerlukan perangkat keras yang kuat. Menurut (Armbrust et al., 2010), komputasi awan menawarkan fleksibilitas, skalabilitas, dan efisiensi biaya, yang sangat bermanfaat bagi peneliti dan pengembang.

### **4) Software Visualisasi Data Dan Evaluasi Model**

Matplotlib dan Seaborn keduanya adalah pustaka Python terkemuka untuk visualisasi data, namun mereka melayani tujuan yang berbeda dan menunjukkan karakteristik yang berbeda. Matplotlib adalah pustaka dasar yang menyediakan kontrol ekstensif atas kustomisasi plot, memungkinkan pengguna untuk membuat berbagai visualisasi statis, animasi, dan interaktif. Namun, seringkali membutuhkan lebih banyak kode dan penyesuaian manual untuk mencapai estetika yang diinginkan (Adebanjo & Banchani, 2023). Sebaliknya, Seaborn dibangun di atas Matplotlib dan dirancang khusus untuk visualisasi data statistik, menawarkan antarmuka tingkat tinggi yang menyederhanakan pembuatan visualisasi kompleks



dengan lebih sedikit kode. Ini secara otomatis menangani transformasi statistik dan terintegrasi dengan mulus dengan struktur data panda, membuatnya sangat efektif untuk analisis data eksplorasi (Waskom, 2021). API deklaratif Seaborn memungkinkan pengguna untuk menghasilkan grafik informatif dengan cepat, sementara juga menyediakan opsi untuk penyesuaian, sehingga melayani pengguna pemula dan berpengalaman (Jose et al., 2023).

## **BAB IV**

### **PELAKSANAAN KERJA PRAKTIK**

#### **IV.1 Input**

Sebagai bagian dari Final Project dalam Kelas *Data Science for Business Development*, setiap mahasiswa diberikan dataset yang berisi informasi mengenai nasabah bank terkait penggunaan layanan kartu kredit untuk di analisis. Tujuan utama analisis ini adalah untuk memahami pola perilaku nasabah yang *churn*, yaitu nasabah yang menutup layanan kartu kredit mereka, serta membangun model prediktif yang mampu mengidentifikasi nasabah dengan risiko *churn* tinggi.

##### **1. Masalah dan Tujuan Proyek**

###### **a. Masalah**

Churn pelanggan kartu kredit merupakan tantangan besar bagi industri perbankan dan keuangan. Dalam dataset Credit Card Churn, churn didefinisikan sebagai nasabah yang berhenti menggunakan kartu kredit mereka atau tidak lagi aktif bertransaksi. Beberapa masalah utama yang muncul dari fenomena churn ini meliputi :

###### **1) Kerugian Finansial bagi Bank**

Kehilangan pelanggan berarti kehilangan sumber pendapatan dari biaya transaksi, bunga kartu kredit, dan layanan tambahan lainnya. Biaya untuk memperoleh pelanggan baru (customer acquisition cost - CAC) seringkali lebih tinggi dibandingkan mempertahankan pelanggan lama (customer retention).

###### **2) Menurunnya Loyalitas Pelanggan**

Pelanggan yang tidak puas dengan layanan kartu kredit bisa berpindah ke bank lain yang menawarkan program yang lebih baik. Persaingan yang semakin ketat di industri perbankan membuat perusahaan harus terus meningkatkan layanan agar tidak kehilangan pelanggan.

###### **3) Perilaku Pelanggan yang Berubah**

Beberapa pelanggan mungkin beralih ke metode pembayaran lain, seperti dompet digital atau sistem kredit alternatif. Ketidakaktifan pelanggan dalam menggunakan kartu kredit bisa menjadi tanda bahwa mereka tidak lagi

membutuhkan layanan tersebut atau memiliki pengalaman negatif dengan bank.

#### 4) Kesulitan dalam Mendeteksi Pelanggan Berisiko Churn

Tanpa analisis yang tepat, sulit bagi bank untuk mengetahui pelanggan mana yang berisiko tinggi berhenti menggunakan kartu kredit mereka. Keputusan strategis dalam mempertahankan pelanggan harus didasarkan pada data dan pola yang jelas.

### **b. Tujuan Proyek**

Proyek ini bertujuan untuk memahami faktor - faktor utama yang menyebabkan churn pelanggan kartu kredit serta mengembangkan strategi untuk mengurangi churn dan meningkatkan retensi pelanggan. Tujuan utama proyek ini meliputi :

#### 1) Analisis Faktor-Faktor yang Berpengaruh terhadap Churn

Mengidentifikasi variabel utama yang berkorelasi dengan churn, seperti lama menjadi nasabah, jumlah transaksi, rasio pemakaian kredit, jumlah kontak ke customer service, dan status sosial - ekonomi pelanggan. Mengetahui apakah pelanggan churn lebih cenderung memiliki karakteristik tertentu, seperti lebih sedikit melakukan transaksi atau sering menghubungi customer service.

#### 2) Pengembangan Model Prediksi Churn

Menggunakan teknik analisis data seperti regresi logistik, decision tree, atau machine learning untuk membangun model yang dapat memprediksi kemungkinan churn pelanggan. Membantu bank dalam mengantisipasi pelanggan yang berisiko tinggi dan mengambil langkah-langkah preventif.

#### 3) Penyusunan Strategi Retensi Pelanggan

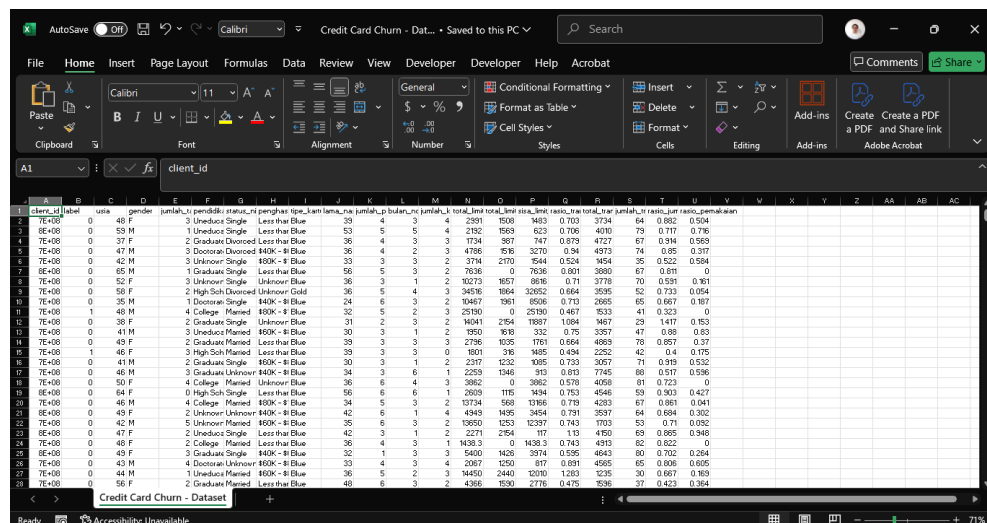
Berdasarkan hasil analisis, merekomendasikan strategi yang dapat diterapkan untuk mempertahankan pelanggan, seperti :

- Meningkatkan layanan pelanggan.
- Menawarkan insentif atau program loyalitas untuk pelanggan berisiko tinggi.
- Memodifikasi kebijakan kredit untuk meningkatkan kepuasan pelanggan.

#### 4) Optimasi Keputusan Bisnis

Membantu bank dalam mengambil keputusan berbasis data untuk mengurangi churn dan meningkatkan profitabilitas. Memungkinkan bank untuk lebih fokus pada segmen pelanggan yang bernilai tinggi dan memiliki potensi retensi yang lebih baik.

## II. Deskripsi Dataset



Gambar 1 Credit Card Churn - Dataset

Dataset ini berisi data nasabah bank yang mencakup tiga aspek utama yaitu Informasi Demografis, Perilaku Penggunaan Kartu Kredit, dan Status churn. Informasi ini dirancang untuk membantu bank mengidentifikasi pola pelanggan yang berisiko churn sehingga dapat dilakukan tindakan pencegahan.

### 1. Informasi Demografis

```
import pandas as pd

# Baca dataset
df = pd.read_csv('/content/Credit Card Churn - Dataset.csv')

# Tentukan kolom yang ingin ditampilkan
kolom_tampilan = ['client_id', 'usia', 'gender', 'jumlah_tanggungan', 'pendidikan', 'status_nikah', 'penghasilan_tahunan']

# Filter data dan tampilkan
info_geografis = df[kolom_tampilan]
print(info_geografis)
```

Gambar 2 Kode untuk menampilkan Informasi Demografis

	client_id	usia	gender	jumlah_tanggungan	pendidikan	status_nikah
0	719455083	48	F	3	Uneducated	Single
1	773503308	59	M	1	Uneducated	Single
2	715452408	37	F	2	Graduate	Divorced
3	711264033	47	M	3	Doctorate	Divorced
4	718943508	42	M	3	Unknown	Single
...	...	...	...	...	...	...
4995	708504783	47	M	3	Uneducated	Unknown
4996	709249083	40	M	2	Unknown	Single
4997	713144208	51	F	0	High School	Married
4998	710375283	26	F	0	Uneducated	Single
4999	719326758	52	F	0	Doctorate	Married

	penghasilan_tahunan
0	Less than \$40K
1	Less than \$40K
2	Less than \$40K
3	\$40K - \$60K
4	\$80K - \$120K
...	...
4995	\$40K - \$60K
4996	\$60K - \$80K
4997	Unknown
4998	\$40K - \$60K
4999	Less than \$40K

[5000 rows x 7 columns]

*Gambar 3 Output Kode untuk menampilkan Informasi Demografis*

Bagian ini mencakup data pribadi nasabah, yang memberikan konteks demografis dan sosial ekonomi :

- 1) client\_id : ID unik yang digunakan untuk mengidentifikasi setiap nasabah.
- 2) Usia : Umur nasabah (dalam tahun), yang memberikan gambaran tentang kelompok umur pelanggan.
- 3) Gender : Jenis kelamin nasabah, dengan kategori M (Male/Laki-laki) dan F (Female/Perempuan).
- 4) jumlah\_tanggungan : Jumlah orang yang menjadi tanggungan finansial nasabah.
- 5) Pendidikan : Tingkat pendidikan nasabah, seperti : High School, College Graduate, atau lainnya.
- 6) status\_nikah : Status pernikahan nasabah, seperti : Married, Single, Divorced, atau Unknown.
- 7) penghasilan\_tahunan : Kategori penghasilan tahunan dalam dollar, yang menunjukkan kemampuan finansial pelanggan.

## 2. Perilaku Penggunaan Kartu Kredit.

Data ini menunjukkan pola interaksi nasabah dengan layanan kartu kredit:

- 1) tipe\_kartu\_kredit : Jenis kartu kredit yang digunakan nasabah, seperti : Blue, Silver, Gold, atau Platinum.

- 2) lama\_nasabah : Durasi (dalam bulan) nasabah telah menjadi pelanggan bank.
- 3) jumlah\_produk : Jumlah produk atau layanan lain dari bank yang dimiliki nasabah.
- 4) bulan\_nonactive : Jumlah bulan dalam 12 bulan terakhir di mana nasabah tidak aktif menggunakan kartu kredit.
- 5) jumlah\_kontak : Jumlah interaksi antara nasabah dan bank dalam 12 bulan terakhir.
- 6) total\_limit\_kredit : Total batas kredit yang diberikan kepada nasabah.
- 7) total\_limit\_kredit\_dipakai : Jumlah kredit yang telah digunakan oleh nasabah.
- 8) sisa\_limit\_kredit : Limit kredit yang masih tersedia.
- 9) rasio\_transaksi\_Q4\_Q1 : Perbandingan nilai total transaksi pada kuartal ke-4 terhadap kuartal ke-1.
- 10) total\_transaksi : Total nilai transaksi nasabah dalam 12 bulan terakhir.
- 11) jumlah\_transaksi : Jumlah transaksi yang dilakukan nasabah dalam 12 bulan terakhir.
- 12) rasio\_jumlah\_transaksi\_Q4\_Q1 : Perbandingan jumlah transaksi pada kuartal ke-4 terhadap kuartal ke-1.
- 13) rasio\_pemakaian : Rasio kredit yang digunakan dibandingkan dengan batas kredit yang diberikan.

### 3. Status Churn

Bagian ini merupakan target variabel yang menjadi fokus analisis :

- 1) Label : Status churn nasabah, di mana 1 menunjukkan nasabah yang churn (menutup layanan kartu kredit) dan 0 menunjukkan nasabah yang tidak churn.

```

import pandas as pd

# Baca dataset
df = pd.read_csv('/content/Credit Card Churn - Dataset.csv')

# Menampilkan status churn
print("Status Churn Nasabah:")
print(df[['client_id', 'label']]) # Menampilkan kolom 'client_id' dan 'label'

# Atau, jika ingin menampilkan keterangan churn:
for index, row in df.iterrows():
    client_id = row['client_id']
    label = row['label']
    status_churn = "Churn" if label == 1 else "Tidak Churn"
    print(f"Client ID: {client_id}, Status: {status_churn}")

```

Gambar 4 Kode untuk menampilkan Status Churn (Label)

```

Output streaming akan dipotong hingga 5000 baris terakhir
Client ID: 719455083, Status: Tidak Churn
Client ID: 773503308, Status: Tidak Churn
Client ID: 715452408, Status: Tidak Churn
Client ID: 711264033, Status: Tidak Churn
Client ID: 718943508, Status: Tidak Churn
Client ID: 778247358, Status: Tidak Churn
Client ID: 710431158, Status: Tidak Churn
Client ID: 715252383, Status: Tidak Churn
Client ID: 717189183, Status: Tidak Churn
Client ID: 712850933, Status: Churn
Client ID: 714426783, Status: Tidak Churn
Client ID: 721001883, Status: Tidak Churn
Client ID: 716352033, Status: Tidak Churn
Client ID: 713696808, Status: Churn
Client ID: 711199908, Status: Tidak Churn
Client ID: 711203658, Status: Tidak Churn
Client ID: 708538608, Status: Tidak Churn
Client ID: 816528333, Status: Tidak Churn
Client ID: 719835033, Status: Tidak Churn
Client ID: 753856008, Status: Tidak Churn
Client ID: 735008958, Status: Tidak Churn
Client ID: 814611633, Status: Tidak Churn
Client ID: 711583458, Status: Tidak Churn
Client ID: 779701983, Status: Tidak Churn
Client ID: 719042208, Status: Tidak Churn
Client ID: 717929133, Status: Tidak Churn
Client ID: 714522258, Status: Tidak Churn
Client ID: 716244933, Status: Tidak Churn
Client ID: 779749908, Status: Tidak Churn
Client ID: 713440833, Status: Tidak Churn
Client ID: 803335008, Status: Tidak Churn
Client ID: 721448283, Status: Churn
Client ID: 713768358, Status: Tidak Churn

```

Gambar 5 Output Kode untuk menampilkan Status Churn (Label)

#### 4. Tujuan Penggunaan Dataset

Dataset ini digunakan untuk Membangun model prediktif menggunakan machine learning untuk mengidentifikasi pelanggan yang berisiko churn, Membantu bank memahami faktor-faktor yang memengaruhi churn, seperti usia, penggunaan kartu kredit, atau jumlah transaksi dan Mendukung bank dalam merancang strategi retensi pelanggan yang efektif.

## 5. Fitur Utama

Fitur utama dari dataset ini adalah variabel - variabel penting yang berkontribusi dalam analisis churn nasabah, meliputi :

- a. Usia (usia) : Variabel ini merepresentasikan umur nasabah dalam tahun. Usia penting untuk mengidentifikasi preferensi perilaku berdasarkan kelompok umur.
- b. Jumlah Tanggungan (jumlah\_tanggungan) : Menunjukkan berapa banyak individu yang menjadi tanggungan finansial nasabah, yang dapat memengaruhi keputusan mereka dalam menggunakan layanan kartu kredit.
- c. Pendidikan (pendidikan) : Tingkat pendidikan nasabah, seperti High School atau College Graduate, yang dapat memengaruhi pola pengelolaan keuangan dan preferensi layanan perbankan.
- d. Penghasilan Tahunan (penghasilan\_tahunan) : Kategori penghasilan tahunan dalam dolar. Variabel ini membantu memahami kemampuan finansial nasabah untuk menggunakan kartu kredit dan kecenderungan untuk churn.
- e. Tipe Kartu Kredit (tipe\_kartu\_kredit) : Jenis kartu kredit yang dimiliki nasabah, seperti Blue, Silver, Gold, atau Platinum, yang mencerminkan segmen pasar atau produk yang digunakan.
- f. Total Transaksi (total\_transaksi) : Total nominal transaksi nasabah dalam 12 bulan terakhir. Variabel ini memberikan gambaran tentang intensitas penggunaan kartu kredit.

## 6. Tujuan Input Data

Dataset ini dirancang untuk digunakan dalam :

- a. Membuat Model Prediktif

Dataset ini digunakan untuk membangun model machine learning yang bertujuan memprediksi nasabah dengan risiko churn tinggi. Variabel target adalah label churn, yang menunjukkan apakah nasabah churn (1) atau tidak churn (0).



b. Identifikasi Nasabah Risiko Tinggi

Dengan menganalisis variabel-variabel penting seperti usia, penghasilan tahunan, dan total transaksi, model dapat mengidentifikasi nasabah yang cenderung berhenti menggunakan layanan kartu kredit.

c. Tindakan Pencegahan oleh Bank

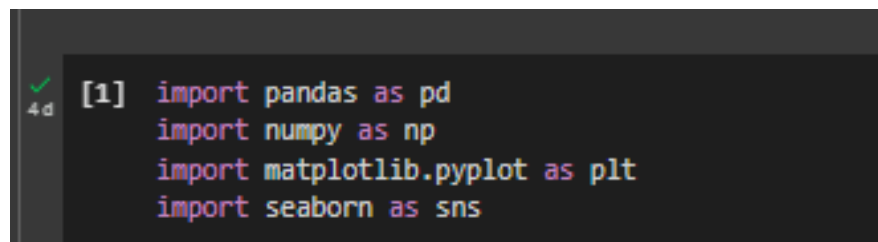
Hasil prediksi dari model dapat digunakan oleh bank untuk merancang strategi retensi pelanggan. Misalnya, bank dapat memberikan promosi khusus, peningkatan layanan, atau konsultasi keuangan kepada nasabah dengan risiko churn tinggi.

## IV.2 Proses

### A. Library

#### 1. Mengimpor Library

Library adalah kumpulan kode yang telah ditulis sebelumnya dan menyediakan fungsi-fungsi siap pakai untuk tugas-tugas tertentu. Terdapat empat library penting yang umum digunakan dalam analisis data dan visualisasi: pandas, numpy, matplotlib, dan seaborn.



```
[1] import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

*Gambar 6 Import Library*

a. Mengimpor Pandas

Pandas adalah library yang sangat berguna untuk manipulasi dan analisis data. Ia menyediakan struktur data yang disebut DataFrame, yang mirip dengan tabel di spreadsheet, dan berbagai fungsi untuk mengolah data dalam DataFrame. Untuk mengimpor pandas, kita gunakan kode : `import pandas as pd`. bahwa `as pd` memberikan alias `pd` untuk library pandas. Ini berarti kita dapat menggunakan `pd` sebagai singkatan untuk mengakses fungsi-fungsi pandas, misalnya `pd.read_csv()` untuk membaca data dari file CSV.

b. Mengimpor NumPy

Selanjutnya, mengimpor library numpy. NumPy adalah library fundamental untuk komputasi numerik di Python. Ia menyediakan struktur data array yang efisien dan berbagai fungsi matematika untuk operasi pada array. Kode untuk mengimpor numpy adalah `import numpy as np`. seperti pandas, `as np` memberikan alias `np` untuk numpy. Kita dapat menggunakan `np` untuk mengakses fungsi-fungsi numpy, misalnya `np.array()` untuk membuat array.

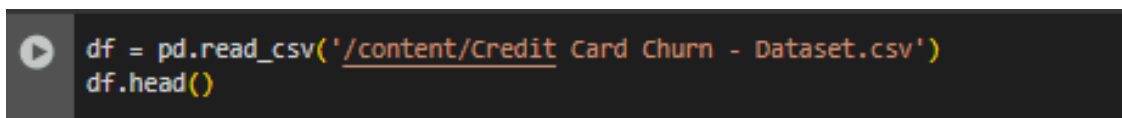
c. Mengimpor Matplotlib

Untuk visualisasi data dengan mengimpor library matplotlib. Matplotlib adalah library yang populer untuk membuat plot dan grafik di Python. Ia menyediakan berbagai jenis plot, seperti plot garis, plot batang, plot scatter, dan banyak lagi. Kode untuk mengimpor pyplot adalah `import matplotlib.pyplot as plt`. Alias `plt` diberikan untuk pyplot, sehingga kita dapat menggunakan `plt` untuk mengakses fungsi-fungsi pyplot, misalnya `plt.plot()` untuk membuat plot garis.

d. Mengimpor Seaborn

Terakhir, mengimpor library seaborn. Seaborn dibangun di atas matplotlib dan menyediakan antarmuka tingkat yang lebih tinggi untuk membuat plot yang lebih informatif dan menarik secara visual. Ia sering digunakan untuk visualisasi data statistik, seperti distribusi, hubungan, dan pola dalam data. Kode untuk mengimpor seaborn adalah `import seaborn as sns`. Alias `sns` diberikan untuk seaborn, sehingga kita dapat menggunakan `sns` untuk mengakses fungsi-fungsi seaborn, misalnya `sns.histplot()` untuk membuat histogram.

## 2. Membaca Dan Menampilkan Data Csv Dengan Pandas



```
df = pd.read_csv('/content/Credit Card Churn - Dataset.csv')
df.head()
```

Gambar 7 Kode untuk membaca dan menampilkan data CSV dengan pandas

a. Membaca Data dari File CSV

Kode Membaca Data dari File CSV adalah

```
df = pd.read_csv('/content/Credit Card Churn - Dataset.csv')
```

Keterangan :

- 1) **pd.read\_csv()** adalah fungsi yang disediakan oleh pandas untuk membaca data dari file CSV.
- 2) **'/content/Credit Card Churn - Dataset.csv'** adalah path atau lokasi file CSV yang ingin kita baca. Pastikan path ini benar agar pandas dapat menemukan file tersebut.
- 3) **df** adalah variabel yang kita gunakan untuk menyimpan data yang telah dibaca. Data tersebut akan disimpan dalam bentuk DataFrame, yaitu struktur data tabular seperti tabel yang disediakan oleh pandas.

b. Menampilkan Data

Kode untuk menampilkan data adalah **df.head()** ini digunakan untuk menampilkan 5 baris pertama dari DataFrame df. Fungsi head() berguna untuk melihat sekilas data yang telah kita baca, termasuk nama kolom dan beberapa nilai data. Dengan menjalankan kode ini, kamu akan melihat output berupa tabel yang berisi 5 baris pertama dari data dalam file "Credit Card Churn - Dataset.csv".

The image shows two screenshots from a Jupyter Notebook. The top screenshot displays the first 5 rows of a dataset with 21 columns. The columns include client\_id, label, usia, gender, jumlah\_tanggungan, pendidikan, status\_nikah, penghasilan\_tahunan, tipe\_kartu\_kredit, lama\_nasabah, bulan\_nonactive, jumlah\_kontak, total\_limit\_kredit, and total\_limit\_kredit\_dipakai. The bottom screenshot shows a subset of columns: bulan\_nonactive, jumlah\_kontak, total\_limit\_kredit, total\_limit\_kredit\_dipakai, sisa\_limit\_kredit, rasio\_transaksi\_Q4\_Q1, total\_transaksi, jumlah\_transaksi, rasio\_jumlah\_transaksi\_Q4\_Q1, and rasio\_pemakaian.

	client_id	label	usia	gender	jumlah_tanggungan	pendidikan	status_nikah	penghasilan_tahunan	tipe_kartu_kredit	lama_nasabah	...	bulan_nonactive	jumlah_kontak	total_limit_kredit	total_limit_kredit_dipakai
0	719455083	0	48	F	3	Uneducated	Single	Less than \$40K	Blue	39	...	3	4	2991.0	1508
1	773503308	0	59	M	1	Uneducated	Single	Less than \$40K	Blue	53	...	5	4	2192.0	1569
2	715452408	0	37	F	2	Graduate	Divorced	Less than \$40K	Blue	36	...	3	3	1734.0	987
3	711264033	0	47	M	3	Dodorate	Divorced	\$40K - \$60K	Blue	36	...	2	3	4786.0	1516
4	718943508	0	42	M	3	Unknown	Single	\$80K - \$120K	Blue	33	...	3	2	3714.0	2170

5 rows x 21 columns

	bulan_nonactive	jumlah_kontak	total_limit_kredit	total_limit_kredit_dipakai	sisa_limit_kredit	rasio_transaksi_Q4_Q1	total_transaksi	jumlah_transaksi	rasio_jumlah_transaksi_Q4_Q1	rasio_pemakaian
3	3	4	2991.0	1508	1483.0	0.703	3734	64	0.882	0.504
5	5	4	2192.0	1569	623.0	0.706	4010	79	0.717	0.716
3	3	3	1734.0	987	747.0	0.879	4727	67	0.914	0.569
2	2	3	4786.0	1516	3270.0	0.940	4973	74	0.850	0.317
3	3	2	3714.0	2170	1544.0	0.524	1454	35	0.522	0.584

Gambar 8 Output Kode untuk membaca dan menampilkan data CSV dengan pandas

## B. Exploratory Data Analysis (EDA)

### 1. Visualisasi Data Kategorikal dengan Seaborn

```
import matplotlib.pyplot as plt
import seaborn as sns

# Data
gender_data = {'F': 2675, 'M': 2325}
pendidikan_data = {'Graduate': 1508, 'High School': 998, 'Uneducated': 755, 'Unknown': 743,
                   'College': 512, 'Post-Graduate': 259, 'Doctorate': 225}
status_nikah_data = {'Married': 2300, 'Single': 1956, 'Unknown': 390, 'Divorced': 354}
penghasilan_data = {'Less than $40K': 1763, '$40K - $60K': 890, '$80K - $120K': 760,
                    '$60K - $80K': 666, 'Unknown': 561, '$120K +': 360}
tipe_kartu_data = {'Blue': 4652, 'Silver': 278, 'Gold': 59, 'Platinum': 11}
label_data = {0: 4200, 1: 800}

# Plotting
plt.figure(figsize=(14, 10))

# Gender
plt.subplot(231)
sns.barplot(x=list(gender_data.keys()), y=list(gender_data.values()))
plt.title('Gender Distribution')

# Pendidikan
plt.subplot(232)
sns.barplot(x=list(pendidikan_data.keys()), y=list(pendidikan_data.values()))
plt.title('Education Level Distribution')
plt.xticks(rotation=45)

# Status Nikah
plt.subplot(233)
sns.barplot(x=list(status_nikah_data.keys()), y=list(status_nikah_data.values()))
plt.title('Marital Status Distribution')

# Penghasilan Tahunan
plt.subplot(234)
sns.barplot(x=list(penghasilan_data.keys()), y=list(penghasilan_data.values()))
plt.title('Annual Income Distribution')
plt.xticks(rotation=45)

# Tipe Kartu Kredit
plt.subplot(235)
sns.barplot(x=list(tipe_kartu_data.keys()), y=list(tipe_kartu_data.values()))
plt.title('Credit Card Type Distribution')

# Label (Churn)
plt.subplot(236)
sns.barplot(x=list(label_data.keys()), y=list(label_data.values()))
plt.title('Churn Label Distribution')

plt.tight_layout()
plt.show()
```

Gambar 9 Kode Visualisasi Data Kategorikal dengan Seaborn

#### a. Persiapan Data

Persiapkan data yang akan divisualisasikan. Data tersebut akan disimpan dalam bentuk dictionary, dimana key adalah kategori dan value adalah frekuensinya. Kode untuk Persiapan Data :

```
gender_data = {'F': 2675, 'M': 2325}
```

```
pendidikan_data = {'Graduate': 1508, 'High School': 998, 'Uneducated': 755,
                   'Unknown': 743, 'College': 512, 'Post-Graduate': 259, 'Doctorate': 225}
```

```
status_nikah_data = {'Married': 2300, 'Single': 1956, 'Unknown': 390,
                    'Divorced': 354}
```

```
# ... (dan dictionary lainnya untuk variabel lain)
```

b. Membuat Figure dan Subplot

Sebelum membuat plot, perlu membuat figure dan subplot. Figure adalah kanvas tempat plot akan digambar, sedangkan subplot adalah bagian-bagian dari figure yang berisi plot individual. Kode untuk Membuat Figure dan Subplot :

```
plt.figure(figsize=(14, 10)) # Membuat figure dengan ukuran 14x10 inci  
plt.subplot(231) # Membuat subplot pertama dalam grid 2x3
```

c. Membuat Bar Plot

Membuat bar plot menggunakan fungsi `sns.barplot()`. Fungsi ini menerima dua argumen utama :

X : daftar kategori yang akan ditampilkan pada sumbu x.

Y : daftar frekuensi yang akan ditampilkan pada sumbu y.

Kode untuk Membuat Bar Plot :

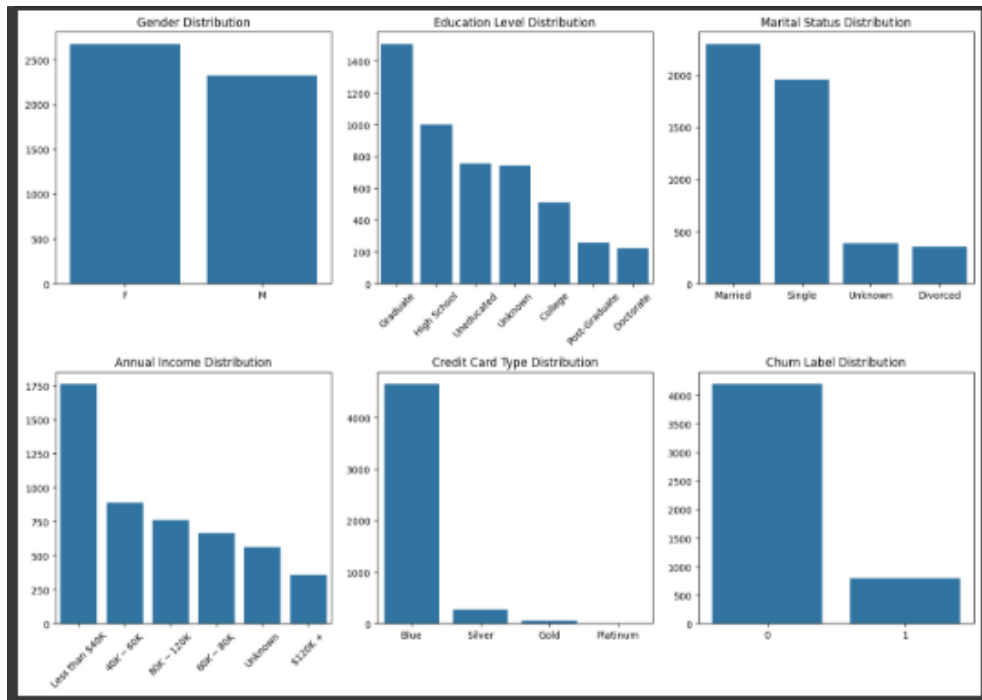
```
sns.barplot(x=list(gender_data.keys()),y=list(gender_data.values()))  
plt.title('Gender Distribution') # Menambahkan judul plot
```

d. Mengulang untuk Variabel Lain

Ulangi langkah Figure dan Subplot dan Bar Plot untuk membuat bar plot untuk setiap variabel kategorikal yang ingin divisualisasikan. Pastikan untuk mengubah nomor subplot (misalnya, `plt.subplot(232)`) dan data yang digunakan.

e. Menampilkan Plot

Setelah semua plot dibuat, gunakan fungsi `plt.tight_layout()` untuk mengatur tata letak subplot agar tidak tumpang tindih. Kemudian, gunakan fungsi `plt.show()` untuk menampilkan figure yang berisi semua plot. Kode untuk menampilkan plot adalah `plt.tight_layout()`, `plt.show()`. Setelah berhasil membuat visualisasi data kategorikal menggunakan Seaborn dapat menyesuaikan plot lebih lanjut, seperti menambahkan label sumbu, mengubah warna, dan sebagainya, dengan memanfaatkan fungsi-fungsi yang disediakan oleh Matplotlib dan Seaborn.



Gambar 10 Output Kode Visualisasi Data Kategorikal dengan Seaborn

## 2. Menggabungkan Data One-Hot Encoded dan Non-One-Hot Encoded dengan Pandas

```
[16] import pandas as pd

# Misalkan df_encoded adalah dataframe hasil one-hot encoding
# dan df_non_one_hot adalah dataframe untuk variabel non-one-hot

# Contoh pembuatan dataframes
data_encoded = {
    'client_id': [1, 2, 3, 4],
    'gender_F': [1, 0, 1, 0],
    'gender_M': [0, 1, 0, 1],
    # tambahkan kolom-kolom one-hot encoded lainnya
}

data_non_one_hot = {
    'client_id': [1, 2, 3, 4],
    'usia': [30, 25, 35, 40],
    'pendidikan': ['Graduate', 'High School', 'Unknown', 'Post-Graduate'],
    # tambahkan kolom-kolom non-one-hot lainnya
}

# Buat dataframe dari data yang ada
df_encoded = pd.DataFrame(data_encoded)
df_non_one_hot = pd.DataFrame(data_non_one_hot)

# Gabungkan dataframe df_encoded dengan df_non_one_hot
# Pastikan untuk menghindari duplikasi kolom 'client_id'
df_merged = pd.merge(df_encoded, df_non_one_hot, on='client_id')

# Tampilkan dataframe hasil penggabungan
print(df_merged.head())
```

Gambar 11 Kode untuk menggabungkan Data One-Hot Encoded dan Non-One Hot Encoded

a. Buat Data Sample

Membuat dua dictionary untuk menyimpan data sample, yang akan mewakili DataFrame dengan data one-hot encoded dan non-one-hot encoded :

```
data_encoded = {
    'client_id': [1, 2, 3, 4],
    'gender_F': [1, 0, 1, 0],
    'gender_M': [0, 1, 0, 1],
    # tambahkan kolom-kolom one-hot encoded lainnya
}
data_non_one_hot = {
    'client_id': [1, 2, 3, 4],
    'usia': [30, 25, 35, 40],
    'pendidikan': ['Graduate', 'High School', 'Unknown', 'Post-Graduate'],
    # tambahkan kolom-kolom non-one-hot lainnya
}
```

Keterangan :

data\_encoded : Berisi data yang sudah di-one-hot encoded. Misalnya, kolom gender\_F dan gender\_M merepresentasikan gender client dengan nilai biner (0 atau 1).

data\_non\_one\_hot: Berisi data yang belum di-one-hot encoded, termasuk variabel kategorikal seperti pendidikan dalam bentuk aslinya.

b. Buat DataFrame dari Data Sample

Ubah dictionary data sample tersebut menjadi DataFrame Pandas:

```
df_encoded = pd.DataFrame(data_encoded)
df_non_one_hot = pd.DataFrame(data_non_one_hot)
```

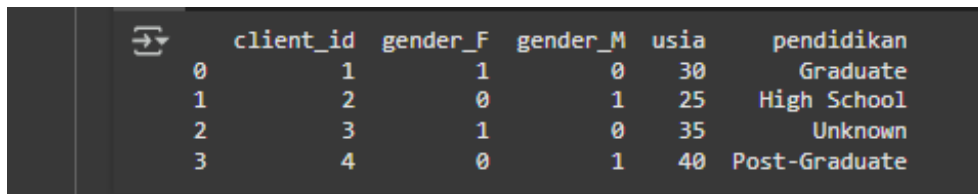
c. Gabungkan DataFrame

Langkah ini adalah inti dari tutorial ini. Gunakan fungsi pd.merge() untuk menggabungkan kedua DataFrame berdasarkan kolom yang sama, yaitu client\_id : df\_merged = pd.merge(df\_encoded, df\_non\_one\_hot, on='client\_id').

on='client\_id' digunakan untuk menentukan kolom kunci yang digunakan untuk menggabungkan DataFrame. Pastikan kolom ini ada di kedua DataFrame.

d. Tampilkan Hasil Penggabungan

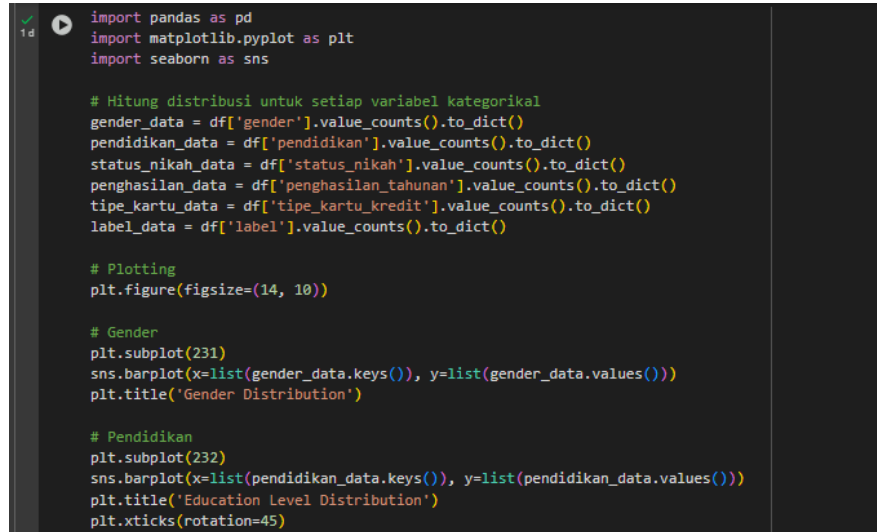
Terakhir, tampilkan beberapa baris pertama dari DataFrame hasil penggabungan untuk memeriksa apakah data sudah tergabung dengan benar : `print(df_merged.head())`. Dengan Langkah - langkah ini, kita berhasil menggabungkan data one-hot encoded dan non-one-hot encoded menjadi satu DataFrame, siap untuk digunakan dalam tahap analisis atau pemodelan selanjutnya.



	client_id	gender_F	gender_M	usia	pendidikan
0	1	1	0	30	Graduate
1	2	0	1	25	High School
2	3	1	0	35	Unknown
3	4	0	1	40	Post-Graduate

Gambar 12 Output Kode untuk menggabungkan Data One-Hot Encoded dan Non-One-Hot Encoded

### 3. Visualisasi Distribusi Variabel Kategorikal dengan Python



```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Hitung distribusi untuk setiap variabel kategorikal
gender_data = df['gender'].value_counts().to_dict()
pendidikan_data = df['pendidikan'].value_counts().to_dict()
status_nikah_data = df['status_nikah'].value_counts().to_dict()
penghasilan_data = df['penghasilan_tahunan'].value_counts().to_dict()
tipe_kartu_data = df['tipe_kartu_kredit'].value_counts().to_dict()
label_data = df['label'].value_counts().to_dict()

# Plotting
plt.figure(figsize=(14, 10))

# Gender
plt.subplot(231)
sns.barplot(x=list(gender_data.keys()), y=list(gender_data.values()))
plt.title('Gender Distribution')

# Pendidikan
plt.subplot(232)
sns.barplot(x=list(pendidikan_data.keys()), y=list(pendidikan_data.values()))
plt.title('Education Level Distribution')
plt.xticks(rotation=45)
```



```

14 [17]
# Status Nikah
plt.subplot(233)
sns.barplot(x=list(status_nikah_data.keys()), y=list(status_nikah_data.values()))
plt.title('Marital Status Distribution')

# Penghasilan Tahunan
plt.subplot(234)
sns.barplot(x=list(penghasilan_data.keys()), y=list(penghasilan_data.values()))
plt.title('Annual Income Distribution')
plt.xticks(rotation=45)

# Tipe Kartu Kredit
plt.subplot(235)
sns.barplot(x=list(tipe_kartu_data.keys()), y=list(tipe_kartu_data.values()))
plt.title('Credit Card Type Distribution')

# Label (Churn)
plt.subplot(236)
sns.barplot(x=list(label_data.keys()), y=list(label_data.values()))
plt.title('Churn Label Distribution')

plt.tight_layout()
plt.show()

```

Gambar 13 Kode Visualisasi Distribusi Variabel Kategorikal dengan Python

a. Muat Dataset

Muat Dataset Anda ke dalam Pandas DataFrame. Gantilah /content/Credit Card Churn - Dataset.csv dengan path ke file dataset Anda. Kode untuk memuat Dataset : `df = pd.read_csv('/content/Credit Card Churn - Dataset.csv')`

Periksa beberapa baris pertama data Anda untuk memahami struktur datanya : `df.head()`

b. Hitung Distribusi

- 1) Gunakan `value_counts()` untuk menghitung frekuensi setiap kategori dalam setiap variabel kategorikal.
- 2) Konversi hasil `value_counts()` ke dalam dictionary menggunakan `to_dict()`.

```

gender_data = df['gender'].value_counts().to_dict()
pendidikan_data = df['pendidikan'].value_counts().to_dict()
# ... (Ulangi untuk variabel kategorikal lainnya) ...

```

c. Visualisasi dengan Bar Chart

- 1) Buat figure dan atur ukurannya menggunakan `plt.figure()` : `plt.figure(figsize=(14, 10))`
- 2) Bagi figure menjadi beberapa subplot menggunakan `plt.subplot()`. Angka dalam `plt.subplot()` menunjukkan tata letak grid (misalnya, 231 berarti grid 2x3 dan subplot pertama).

- 3) Gunakan `sns.barplot()` untuk membuat bar chart untuk setiap variabel.  
x: Daftar kategori (kunci dari dictionary data).  
y: Daftar frekuensi (nilai dari dictionary data).
- 4) Berikan judul untuk setiap plot menggunakan `plt.title()`.
- 5) Rotasi label sumbu x jika diperlukan menggunakan `plt.xticks(rotation=45)`.

```
plt.subplot(231)
```

```
sns.barplot(x=list(gender_data.keys()), y=list(gender_data.values()))
```

```
plt.title('Gender Distribution')
```

```
plt.subplot(232)
```

```
sns.barplot(x=list(pendidikan_data.keys()),
```

```
y=list(pendidikan_data.values()))
```

```
plt.title('Education Level Distribution')
```

```
plt.xticks(rotation=45)
```

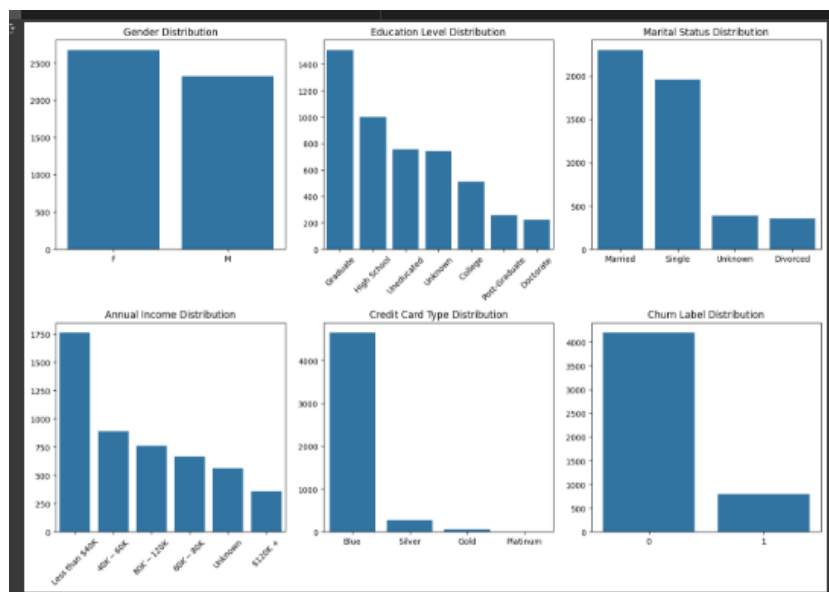
```
# ... (Ulangi untuk variabel kategorikal lainnya) ...
```

- 6) Atur tata letak subplot agar tidak tumpang tindih menggunakan `plt.tight_layout()`.

- 7) Tampilkan plot menggunakan `plt.show()`.

```
plt.tight_layout()
```

```
plt.show()
```



Gambar 14 Output Visualisasi Distribusi Variabel Kategorikal dengan Python

## C. Data Preprocessing

### 1. *One-Hot Encoding* untuk Data Kategorikal dalam *Machine Learning*

One-hot encoding adalah teknik penting dalam machine learning untuk mengubah data kategorikal (misalnya, warna, jenis kelamin, pendidikan) menjadi format numerik yang dapat dipahami oleh algoritma machine learning. Langkah - Langkah melakukan *one-hot encoding* menggunakan Python dan library scikit-learn :

```
# Data preprocessing
# 1. Handling missing value
# Berdasarkan data profile diatas, bahwa tidak ada missing value, maka proses ini akan kita lewati

# 2. Penyandian variabel kategorikal
# Berdasarkan temuan adanya distribusi variabel categorical, maka kita akan lakukan one-hot encode

import pandas as pd
from sklearn.preprocessing import OneHotEncoder

# Baca dataset
df = pd.read_csv('/content/Credit Card Churn - Dataset.csv')

# Menentukan nilai unik dari setiap variabel kategorikal
categorical_columns = ['gender', 'pendidikan', 'status_nikah', 'penghasilan_tahunan', 'tipe_kartu_kredit']

for column in categorical_columns:
    print(f"Unique values in '{column}':")
    print(df[column].unique())
    print()

# Mengonversi variabel kategorikal dengan One-Hot Encoding
# The 'sparse' argument has been deprecated and replaced with 'sparse_output'
one_hot_encoder = OneHotEncoder(sparse_output=False, drop='first') # Changed 'sparse' to 'sparse_output'

# Fit dan transform data kategorikal
encoded_data = one_hot_encoder.fit_transform(df[categorical_columns])

# Mendapatkan nama kolom hasil one-hot encoding
encoded_columns = one_hot_encoder.get_feature_names_out(categorical_columns)

# Membuat DataFrame dari hasil one-hot encoding
df_encoded = pd.DataFrame(encoded_data, columns=encoded_columns)

# Menggabungkan hasil one-hot encoding dengan kolom lain yang tidak di-encode
df_non_encoded = df.drop(columns=categorical_columns)
df_final = pd.concat([df_non_encoded.reset_index(drop=True), df_encoded.reset_index(drop=True)], axis=1)

# Menampilkan beberapa baris hasil penggabungan
print(df_final.head())

# Menyimpan hasil ke file baru jika diperlukan
df_final.to_csv('creditcardchurn_encoded.csv', index=False)

df_final.info()
```

Gambar 15 kode one-hot encoding menggunakan Python dan library scikit-learn

#### a. Persiapan

- 1) Import library: Pastikan Anda telah menginstal library pandas dan scikit-learn. Import library yang dibutuhkan :

```
import pandas as pd
from sklearn.preprocessing import OneHotEncoder
```

- 2) Muat data: Muat dataset yang berisi variabel kategorikal yang ingin Anda encode. Misalnya, kita menggunakan dataset "Credit Card Churn" yang disebutkan sebelumnya :

```
df = pd.read_csv('/content/Credit Card Churn - Dataset.csv')
```

b. Identifikasi Variabel Kategorikal

Tentukan kolom-kolom dalam dataset yang berisi data kategorikal. Dalam contoh ini, kita punya :

```
categorical_columns = ['gender', 'pendidikan', 'status_nikah',  
'penghasilan_tahunan', 'tipe_kartu_kredit']
```

c. One-Hot Encoding

1) Buat objek OneHotEncoder :

```
one_hot_encoder = OneHotEncoder(sparse_output=False, drop='first')
```

a) `sparse_output=False`: Menghasilkan output berupa array NumPy biasa, bukan sparse matrix.

b) `drop='first'`: Menghindari dummy variable trap dengan menghapus kategori pertama dari setiap fitur.

2) Lakukan fit dan transform :

a) `fit`: Objek OneHotEncoder akan mempelajari kategori-kategori yang ada dalam setiap kolom kategorikal.

b) `transform`: Data kategorikal akan diubah menjadi format one-hot encoded.

```
encoded_data =  
one_hot_encoder.fit_transform(df[categorical_columns])
```

3) Dapatkan nama kolom baru: Dapatkan nama kolom hasil one-hot encoding:

```
encoded_columns =  
one_hot_encoder.get_feature_names_out(categorical_columns)
```

4) Buat DataFrame baru: Buat DataFrame baru dari data yang telah di-encode:

```
df_encoded = pd.DataFrame(encoded_data, columns=encoded_columns)
```

d. Gabungkan Data

Gabungkan DataFrame hasil one-hot encoding (`df_encoded`) dengan kolom-kolom lain yang tidak di-encode dari dataset asli :

```
df_non_encoded = df.drop(columns=categorical_columns)  
df_final = pd.concat([df_non_encoded.reset_index(drop=True),  
df_encoded.reset_index(drop=True)], axis=1)
```

e. Simpan dan Gunakan Data

- 1) Simpan data: Simpan DataFrame hasil preprocessing ke file CSV baru: `df_final.to_csv('creditcardchurn_encoded.csv', index=False)`
- 2) Gunakan data: Sekarang, `df_final` berisi data yang telah di-preprocess dan siap digunakan untuk melatih model machine learning Anda.

```

Unique values in 'gender':
['F' 'M']

Unique values in 'pendidikan':
['Uneducated' 'Graduate' 'Doctorate' 'Unknown' 'High School' 'College'
 'Post-Graduate']

Unique values in 'status_nikah':
['Single' 'Divorced' 'Married' 'Unknown']

Unique values in 'penghasilan_tahunan':
['Less than $40K' '$40K - $50K' '$50K - $60K' '$60K - $70K' '$70K - $80K'
 '$80K - $90K' '$90K - $100K' '$100K - $110K' '$110K - $120K' '$120K +']

Unique values in 'tipe_kartu_kredit':
['Blue' 'Gold' 'Silver' 'Platinum']

  client_id  label  usia  jumlah_tanggungan  lama_nasabah  jumlah_produk \
0  719455083    0    48                3           39         4
1  773503308    0    59                1           53         5
2  713423408    0    37                2           36         4
3  711264033    0    47                3           36         4
4  718943508    0    42                3           33         3

  bulan_nonactive  jumlah_kontak  total_limit_kredit \
0                3             4          2991.0
1                5             4          2192.0
2                3             3          1734.0
3                2             3          4786.0
4                3             2          3714.0

  total_limit_kredit_dipakai  ...  status_nikah_Single  status_nikah_Unknown \
0                1598  ...              1.0              0.0
1                1560  ...              1.0              0.0
2                987  ...              0.0              0.0
3                1516  ...              0.0              0.0
4                2178  ...              1.0              0.0

  penghasilan_tahunan_40K - 50K  penghasilan_tahunan_50K - 60K \
0                0.0                0.0
1                0.0                0.0
2                0.0                0.0
3                1.0                0.0
4                0.0                0.0

  penghasilan_tahunan_80K - 90K  penghasilan_tahunan_Less than 40K \
0                0.0                1.0
1                0.0                1.0
2                0.0                1.0
3                0.0                0.0
4                1.0                0.0

  penghasilan_tahunan_Unknown  tipe_kartu_kredit_Gold \
0                0.0                0.0
1                0.0                0.0
2                0.0                0.0
3                0.0                0.0
4                0.0                0.0

  tipe_kartu_kredit_Platinum  tipe_kartu_kredit_Silver
0                0.0                0.0
1                0.0                0.0
2                0.0                0.0
3                0.0                0.0
4                0.0                0.0

[5 rows x 34 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 34 columns):
 #   Column                Non-Null Count  Dtype
---  -
 0   client_id             5000 non-null   int64
 1   label                 5000 non-null   int64
 2   usia                  5000 non-null   int64
 3   jumlah_tanggungan     5000 non-null   int64
 4   lama_nasabah          5000 non-null   int64
 5   jumlah_produk         5000 non-null   int64
 6   bulan_nonactive       5000 non-null   int64
 7   jumlah_kontak         5000 non-null   int64
 8   total_limit_kredit    5000 non-null   float64
 9   total_limit_kredit_dipakai  5000 non-null   int64
10  sisa_limit_kredit      5000 non-null   float64
11  rasio_transaksi_Q4_Q1  5000 non-null   float64
12  total_transaksi       5000 non-null   int64
13  jumlah_transaksi      5000 non-null   int64
14  rasio_jumlah_transaksi_Q4_Q1  5000 non-null   float64
15  rasio_pemakaian       5000 non-null   float64
16  gender_M              5000 non-null   float64
17  pendidikan_Doctorate  5000 non-null   float64
18  pendidikan_Graduate   5000 non-null   float64
19  pendidikan_High_School  5000 non-null   float64
20  pendidikan_Post-Graduate  5000 non-null   float64
21  pendidikan_Uneducated  5000 non-null   float64
22  pendidikan_Unknown    5000 non-null   float64
23  status_nikah_Married  5000 non-null   float64

```

```

24 status_nikah_Single      5000 non-null float64
25 status_nikah_Unknown    5000 non-null float64
26 penghasilan_tahunan_40K - $60K  5000 non-null float64
27 penghasilan_tahunan_60K - $80K  5000 non-null float64
28 penghasilan_tahunan_80K - $120K 5000 non-null float64
29 penghasilan_tahunan_Less than $40K 5000 non-null float64
30 penghasilan_tahunan_Unknown    5000 non-null float64
31 tipe_kartu_kredit_Gold        5000 non-null float64
32 tipe_kartu_kredit_Platinum    5000 non-null float64
33 tipe_kartu_kredit_Silver      5000 non-null float64
dtypes: float64(23), int64(11)
memory usage: 1.3 MB

```

Gambar 16 Output one-hot encoding menggunakan Python dan library scikit-learn

## 2. Memahami Korelasi Data dengan Python

tujuan untuk memahami hubungan antar variabel dalam dataset menggunakan visualisasi korelasi. Berikut adalah Langkah - langkahnya :

```

[19] import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import OneHotEncoder

# Baca dataset
df = pd.read_csv('/content/Credit Card Churn - Dataset.csv', delimiter='\t')

# Menampilkan beberapa baris hasil penggabungan
print(df_final.head())

# Menyimpan hasil ke file baru jika diperlukan
df_final.to_csv('creditcardchurn_encoded.csv', index=False)

# Menghitung matriks korelasi
correlation_matrix = df_final.corr()

# Menampilkan matriks korelasi
print(correlation_matrix)

# Plotting heatmap untuk visualisasi korelasi
plt.figure(figsize=(16, 12))
sns.heatmap(correlation_matrix, annot=True, fmt=".2f", cmap='coolwarm', cbar=True)
plt.title('Correlation Matrix Heatmap')
plt.show()

```

Gambar 17 Kode untuk memahami Korelasi Data dengan Python

### a. Persiapan

Pertama, kita perlu mengimport library yang dibutuhkan :

```
import pandas as pd
```

```
import seaborn as sns
```

```
import matplotlib.pyplot as plt
```

pandas digunakan untuk membaca dan memanipulasi data. seaborn dan matplotlib digunakan untuk visualisasi data.

### b. Membaca Data

Selanjutnya, kita baca dataset yang akan dianalisis menggunakan pandas :

```
df = pd.read_csv('/content/Credit Card Churn - Dataset.csv', delimiter='\t')
```

Kode ini membaca data dari file CSV bernama 'Credit Card Churn - Dataset.csv' dan menyimpannya dalam variabel df sebagai DataFrame. Asumsikan file tersebut menggunakan tab (\t) sebagai pemisah antar kolom.

c. Menghitung Matriks Korelasi

Setelah data terbaca, kita hitung matriks korelasi menggunakan `df.corr()`:

```
correlation_matrix = df_final.corr()
```

Metode `corr()` menghitung korelasi Pearson antara setiap pasangan kolom numerik dalam DataFrame `df_final`. Hasilnya disimpan dalam variabel `correlation_matrix`.

d. Visualisasi dengan Heatmap

Langkah terakhir adalah memvisualisasikan matriks korelasi menggunakan heatmap dengan `seaborn` dan `matplotlib` :

```
plt.figure(figsize=(16, 12))
```

```
sns.heatmap(correlation_matrix, annot=True, fmt=".2f", cmap='coolwarm',  
cbar=True)
```

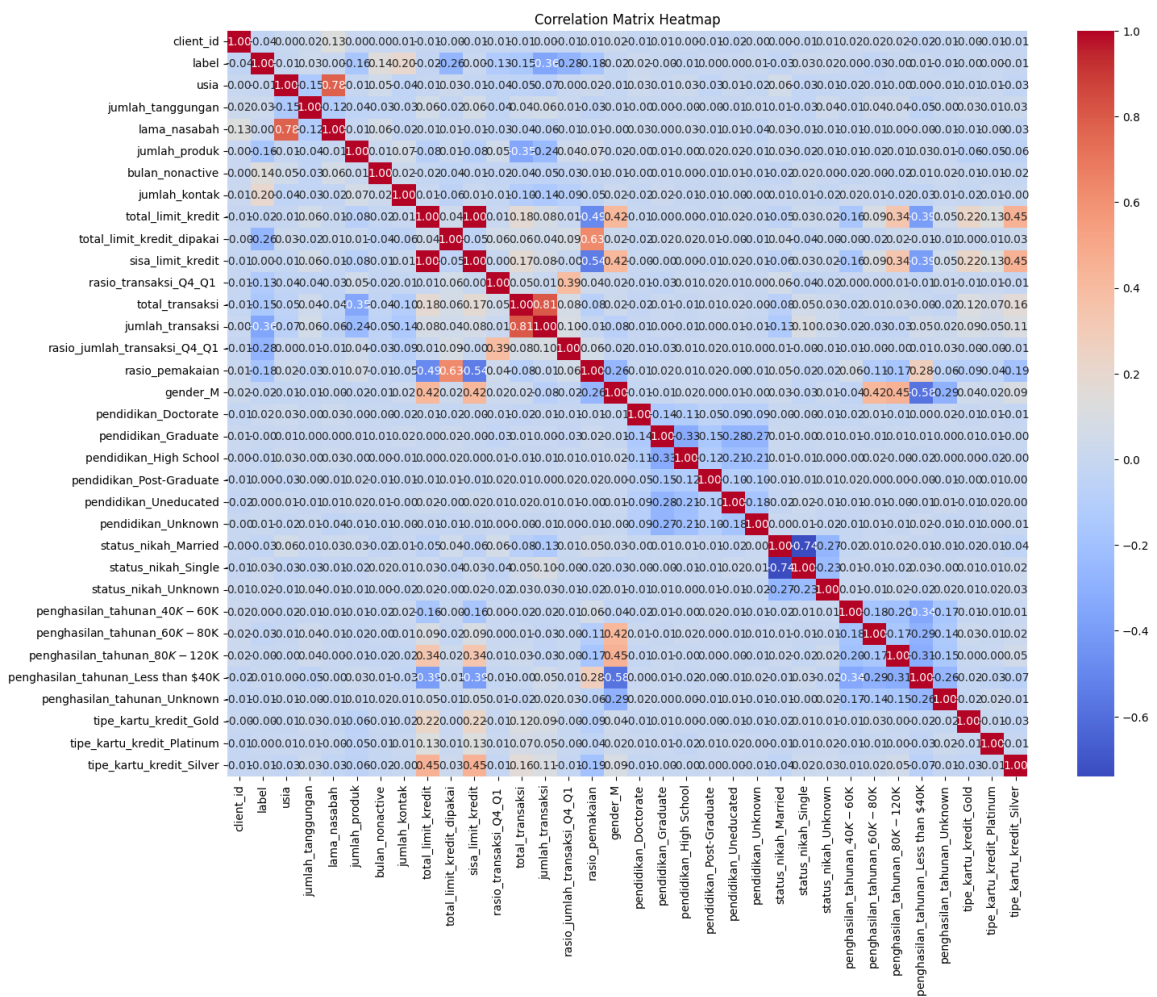
```
plt.title('Correlation Matrix Heatmap')
```

```
plt.show()
```

- 1) `plt.figure(figsize=(16, 12))` mengatur ukuran gambar.
- 2) `sns.heatmap(...)` membuat heatmap dengan parameter:
  - a) `correlation_matrix`: data yang akan divisualisasikan.
  - b) `annot=True`: menampilkan nilai korelasi pada setiap sel.
  - c) `fmt=".2f"`: format nilai korelasi dengan 2 angka di belakang koma.
  - d) `cmap='coolwarm'`: menggunakan skema warna 'coolwarm'.
  - e) `cbar=True`: menampilkan colorbar.
- 3) `plt.title(...)` memberi judul pada heatmap.
- 4) `plt.show()` menampilkan heatmap.

Interpretasi Heatmap :

- a) Warna pada heatmap menunjukkan kekuatan dan arah korelasi.
- b) Warna merah menunjukkan korelasi positif, biru menunjukkan korelasi negatif, dan warna semakin terang menunjukkan korelasi semakin lemah.
- c) Nilai korelasi mendekati 1 atau -1 menunjukkan korelasi yang kuat, sedangkan nilai mendekati 0 menunjukkan korelasi yang lemah.



Gambar 18 Correlation Matrix Heatmap

Deskripsi Gambar 18 Correlation Matrix Heatmap.

## 1. Struktur Heatmap

### a. Sumbu X dan Y

- Menampilkan daftar fitur yang ada dalam dataset.
- Fitur-fitur ini mencakup informasi demografis, riwayat transaksi, pendidikan, status pernikahan, penghasilan tahunan, dan jenis kartu kredit.

### b. Warna dalam Heatmap

- Merah tua : Korelasi positif tinggi (mendekati +1), artinya kedua variabel memiliki hubungan yang kuat dan bergerak searah.
- Biru tua : Korelasi negatif tinggi (mendekati -1), artinya jika satu variabel meningkat, yang lain cenderung menurun.



- Putih/abu-abu : Korelasi rendah atau tidak signifikan (mendekati 0), menunjukkan tidak ada hubungan yang kuat antara variabel.

## 2. Analisis Korelasi

### a. Korelasi antara fitur dan label

#### 1) Fitur dengan korelasi negatif terhadap label

- `total_limit_kredit_dipakai` (-0.26) : Semakin tinggi penggunaan limit kredit, semakin rendah nilai label.
- `jumlah_transaksi` (-0.36) : Semakin banyak transaksi, semakin kecil nilai label.
- `rasio_jumlah_transaksi_Q4_Q1` (-0.28) : Perubahan rasio transaksi antar kuartal juga berpengaruh terhadap label.

#### 2) Fitur dengan korelasi positif terhadap label

- `jumlah_kontak` (0.20) : Jumlah kontak pelanggan memiliki sedikit korelasi positif terhadap label.

#### 3) Korelasi antar fitur penting

- `total_limit_kredit_dipakai` memiliki korelasi positif tinggi (0.63) dengan `rasio_pemakaian`, yang berarti semakin tinggi limit kredit yang digunakan, semakin tinggi rasio penggunaannya.
- `jumlah_transaksi` berkorelasi positif dengan `total_transaksi`, yang menunjukkan bahwa semakin banyak transaksi dilakukan, semakin tinggi total jumlah transaksi.
- Penghasilan tahunan memiliki beberapa korelasi dengan status pernikahan dan jenis kartu kredit, yang bisa menjadi faktor dalam analisis keuangan pelanggan.

## 3. Kesimpulan

- Fitur `jumlah_transaksi`, `rasio_jumlah_transaksi_Q4_Q1`, dan `total_limit_kredit_dipakai` memiliki korelasi negatif terhadap label, yang menunjukkan bahwa faktor ini dapat menjadi indikator dalam prediksi model.
- Fitur seperti `jumlah_kontak` memiliki sedikit korelasi positif, tetapi lebih lemah dibandingkan fitur lain.

- Hubungan antar fitur tertentu, seperti rasio pemakaian dengan total limit kredit, menunjukkan adanya keterkaitan dalam pola penggunaan kredit pelanggan.

### 3. Mencari dan Visualisasi 5 Fitur Terkorelasi dengan Label

Tujuan untuk membantu memahami cara mencari dan memvisualisasikan 5 fitur yang memiliki korelasi paling kuat dengan kolom 'label' (misalnya, customer churn) dalam dataset dengan menggunakan Python, library Pandas, Seaborn, dan Matplotlib untuk mencapai tujuan ini.

```
[20] # ambil 5 variabel dengan nilai korelasi paling tinggi

# Menghitung matriks korelasi
correlation_matrix = df_final.corr()

# Menampilkan matriks korelasi
print(correlation_matrix)

# Mengambil 5 nilai korelasi terkuat yang mempengaruhi label
label_correlation = correlation_matrix['label'].drop('label')
top_5_features = label_correlation.abs().nlargest(5).index

# Membuat DataFrame baru untuk korelasi antara label dan 5 fitur teratas
top_5_correlation_matrix = df_final[top_5_features.tolist() + ['label']].corr()

# Plotting heatmap untuk visualisasi korelasi
plt.figure(figsize=(10, 8))
sns.heatmap(top_5_correlation_matrix, annot=True, fmt=".2f", cmap='coolwarm', cbar=True)
plt.title('Top 5 Feature Correlations with Label')
plt.show()
```

Gambar 19 Mencari dan Visualisasi 5 Fitur Terkorelasi dengan Label

#### a. Import Library

Pertama, kita perlu mengimport library yang dibutuhkan :

```
import pandas as pd
```

```
import seaborn as sns
```

```
import matplotlib.pyplot as plt
```

Pandas : digunakan untuk manipulasi data dalam bentuk DataFrame.

Seaborn: digunakan untuk visualisasi data yang lebih menarik.

Matplotlib: digunakan untuk membuat plot dan grafik.

#### b. Hitung Matriks Korelasi

Selanjutnya, hitung matriks korelasi untuk semua fitur numerik dalam dataset.

Asumsikan dataset kamu sudah tersimpan dalam DataFrame bernama df\_final:

```
correlation_matrix = df_final.corr()
```

```
print(correlation_matrix)
```

`df_final.corr()`: fungsi ini menghitung korelasi antara semua kolom numerik dalam `df_final` dan mengembalikan matriks korelasi.

`print(correlation_matrix)`: menampilkan matriks korelasi di console.

c. Temukan 5 Fitur Terkorelasi

Sekarang, akan mencari 5 fitur yang memiliki korelasi absolut tertinggi dengan kolom 'label' :

```
label_correlation = correlation_matrix['label'].drop('label')
```

```
top_5_features = label_correlation.abs().nlargest(5).index
```

`correlation_matrix['label']`: memilih kolom 'label' dari matriks korelasi, yang berisi korelasi 'label' dengan semua fitur lainnya. `.drop('label')`: menghapus korelasi 'label' dengan dirinya sendiri (yang selalu 1). `.abs()`: mengambil nilai absolut dari korelasi, sehingga kita fokus pada kekuatan korelasi, bukan arahnya. `.nlargest(5)`: memilih 5 fitur dengan nilai korelasi absolut tertinggi. `.index`: mengambil nama-nama dari 5 fitur tersebut.

d. Buat Matriks Korelasi untuk 5 Fitur Teratas

Membuat matriks korelasi baru yang hanya berisi 5 fitur teratas dan kolom 'label' :

```
top_5_correlation_matrix = df_final[top_5_features.tolist() + ['label']].corr()
```

`top_5_features.tolist() + ['label']` : membuat list yang berisi nama 5 fitur teratas dan kolom 'label'. `df_final[...]` : memilih kolom-kolom tersebut dari DataFrame `df_final`. `.corr()` : menghitung matriks korelasi untuk kolom-kolom yang dipilih.

e. Visualisasi dengan Heatmap

Terakhir, visualisasikan matriks korelasi menggunakan heatmap :

```
plt.figure(figsize=(10, 8))
```

```
sns.heatmap(top_5_correlation_matrix, annot=True, fmt=".2f",  
cmap='coolwarm', cbar=True)
```

```
plt.title('Top 5 Feature Correlations with Label')
```

```
plt.show()
```

`sns.heatmap(...)` : membuat heatmap dari `top_5_correlation_matrix`.

`annot=True` : menampilkan nilai korelasi pada heatmap.

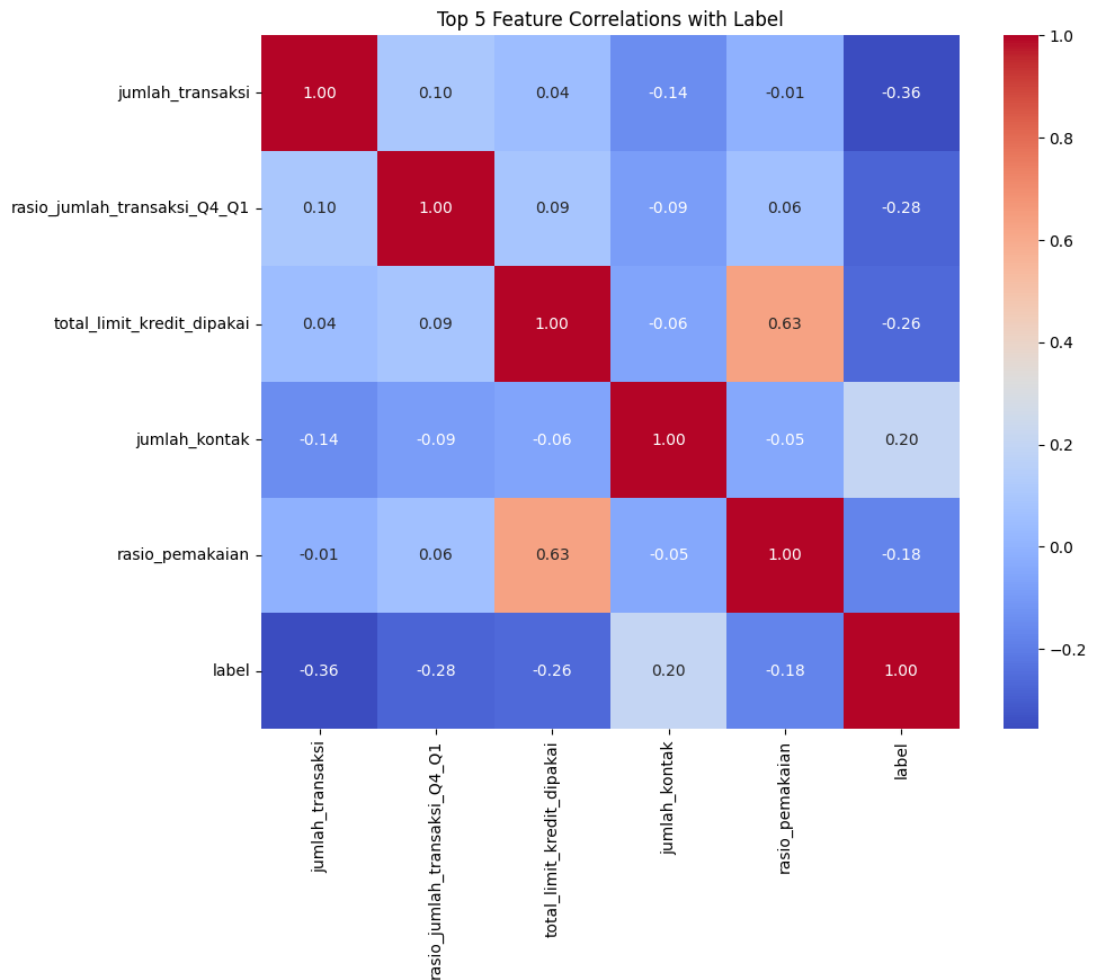
`fmt=".2f"` : memformat nilai korelasi menjadi dua desimal.

`cmap='coolwarm'` : menggunakan skema warna coolwarm (merah untuk korelasi tinggi, biru untuk korelasi rendah).

`cbar=True` : menampilkan color bar untuk interpretasi nilai korelasi.

`plt.title(...)` : memberi judul heatmap.

`plt.show()` : menampilkan heatmap.



Gambar 20 Top 5 Feature Correlations with label

Deskripsi Gambar 20 Top 5 Feature Correlations with label.

### 1. Struktur Heatmap

#### a. Sumbu X dan Y

- Berisi lima fitur yang memiliki korelasi tertinggi dengan label serta hubungan antar fitur.

- Fitur yang ditampilkan adalah :

1. jumlah\_transaksi
2. rasio\_jumlah\_transaksi\_Q4\_Q1
3. total\_limit\_kredit\_dipakai
4. jumlah\_kontak
5. rasio\_pemakaian
6. label

b. Warna dalam Heatmap

- Merah tua : Korelasi positif tinggi (mendekati +1), artinya kedua variabel bergerak searah.
- Biru tua : Korelasi negatif tinggi (mendekati -1), artinya jika satu variabel meningkat, yang lain cenderung menurun.
- Putih/abu-abu : Korelasi rendah atau tidak signifikan (mendekati 0), artinya tidak ada hubungan yang jelas.

## 2. Analisis Korelasi

a. Korelasi antara fitur dan label

- jumlah\_transaksi memiliki korelasi negatif dengan label (-0.36), menunjukkan bahwa semakin banyak transaksi, semakin rendah kemungkinan hasil yang diwakili oleh label.
- rasio\_jumlah\_transaksi\_Q4\_Q1 juga berkorelasi negatif (-0.28), menandakan bahwa perubahan rasio transaksi antar kuartal bisa memengaruhi label dengan pola yang serupa.
- total\_limit\_kredit\_dipakai berkorelasi negatif (-0.26) dengan label, artinya semakin tinggi penggunaan limit kredit, semakin rendah nilai label.
- jumlah\_kontak memiliki korelasi positif kecil (0.20) terhadap label, menunjukkan sedikit hubungan positif.
- rasio\_pemakaian memiliki korelasi negatif lebih kecil (-0.18), menunjukkan pengaruh yang lebih lemah terhadap label.

b. Korelasi antar fitur

- total\_limit\_kredit\_dipakai memiliki korelasi positif sedang (0.63) dengan rasio\_pemakaian, yang berarti semakin tinggi limit kredit yang digunakan, semakin tinggi rasio penggunaannya.
- Sebagian besar fitur memiliki korelasi rendah satu sama lain, menunjukkan bahwa fitur-fitur ini relatif independen dalam hubungan mereka terhadap label.

3. Kesimpulan

- Fitur dengan korelasi negatif lebih tinggi terhadap label (seperti jumlah\_transaksi, rasio\_jumlah\_transaksi\_Q4\_Q1, dan total\_limit\_kredit\_dipakai) dapat menjadi prediktor penting.
- jumlah\_kontak adalah satu-satunya fitur yang memiliki korelasi positif dengan label, meskipun cukup rendah (0.20).
- Korelasi antar fitur menunjukkan bahwa total\_limit\_kredit\_dipakai dan rasio\_pemakaian berkaitan erat, yang dapat memberikan wawasan tambahan dalam analisis lebih lanjut.

4. Memprediksi "Label" dengan Logistic Regression dan Random Forest

Proses membangun dan mengevaluasi dua model machine learning, yaitu Logistic Regression dan Random Forest, untuk memprediksi variabel "label" berdasarkan 5 fitur teratas yang paling berkorelasi dengannya.

```
[21] import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score

# Mengambil 5 fitur teratas yang memiliki korelasi tertinggi dengan label
top_5_features = label.corr().abs().nlargest(5).index
X = df_final[top_5_features]
y = df_final['label']

# Memisahkan data menjadi training dan testing set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Model 1: Logistic Regression
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
y_pred_log_reg = log_reg.predict(X_test)

print("\nLogistic Regression Model")
print("Accuracy:", accuracy_score(y_test, y_pred_log_reg))
print(classification_report(y_test, y_pred_log_reg))

# Model 2: Random Forest
random_forest = RandomForestClassifier(n_estimators=100, random_state=42)
random_forest.fit(X_train, y_train)
y_pred_rf = random_forest.predict(X_test)

print("\nRandom Forest Model")
print("Accuracy:", accuracy_score(y_test, y_pred_rf))
print(classification_report(y_test, y_pred_rf))
```

Gambar 21 Kode Memprediksi "Label" dengan Logistic Regression dan Random Forest

a. Import Library yang Diperlukan

Pertama, kita perlu mengimpor library yang dibutuhkan untuk menjalankan kode :

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score
```

pandas : Digunakan untuk manipulasi dan analisis data. train\_test\_split : Digunakan untuk membagi data menjadi set pelatihan dan pengujian. LogisticRegression : Digunakan untuk membuat model Logistic Regression. RandomForestClassifier: Digunakan untuk membuat model Random Forest. classification\_report, accuracy\_score : Digunakan untuk mengevaluasi kinerja model.

b. Persiapkan Data

Selanjutnya, kita perlu menyiapkan data untuk model :

```
# Mengambil 5 fitur teratas yang memiliki korelasi tertinggi dengan label
top_5_features = label_correlation.abs().nlargest(5).index
X = df_final[top_5_features]
y = df_final['label']
# Memisahkan data menjadi training dan testing set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

- Memilih Fitur : Kode ini memilih 5 fitur teratas yang paling berkorelasi dengan label. Fitur-fitur ini akan digunakan sebagai input untuk model.
- Membuat Variabel Input dan Target : X berisi fitur-fitur yang dipilih, dan y berisi variabel target (label).
- Membagi Data : Data dibagi menjadi set pelatihan (X\_train, y\_train) dan set pengujian (X\_test, y\_test). 80% data digunakan untuk pelatihan dan 20% untuk pengujian. random\_state=42 memastikan pembagian data yang konsisten.

c. Bangun dan Evaluasi Model

Membangun dan mengevaluasi dua model :

**Model 1: Logistic Regression**

```
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
y_pred_log_reg = log_reg.predict(X_test)
print("Logistic Regression Model")
print("Accuracy:", accuracy_score(y_test, y_pred_log_reg))
print(classification_report(y_test, y_pred_log_reg))
```

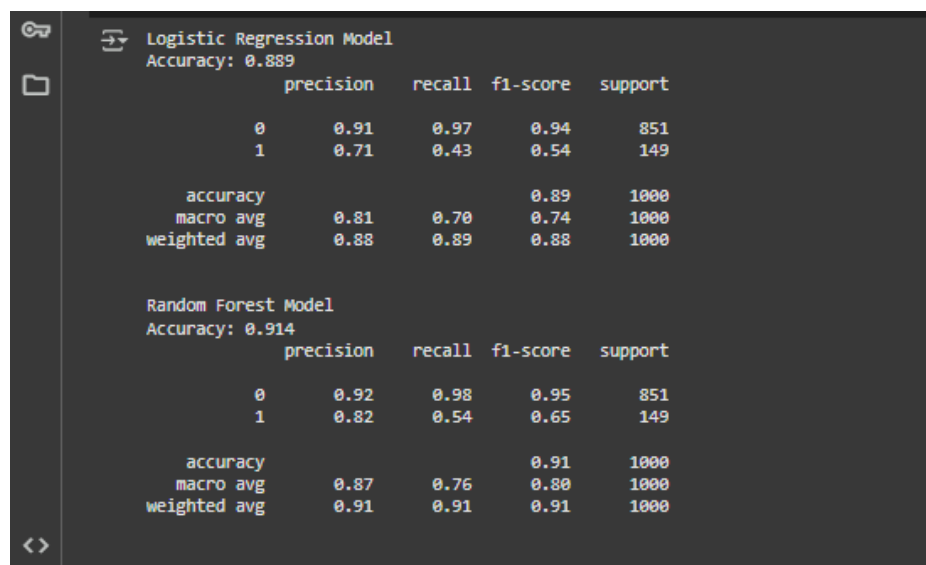
- Membuat Model : Sebuah model Logistic Regression dibuat dan disimpan dalam variabel log\_reg.
- Melatih Model : Model dilatih menggunakan data pelatihan (X\_train, y\_train).
- Membuat Prediksi : Model digunakan untuk membuat prediksi pada data pengujian (X\_test), dan prediksi disimpan dalam y\_pred\_log\_reg.
- Mengevaluasi Model : Kinerja model dievaluasi menggunakan accuracy\_score dan classification\_report.

**Model 2: Random Forest**

```
random_forest = RandomForestClassifier(n_estimators=100,
random_state=42)
random_forest.fit(X_train, y_train)
y_pred_rf = random_forest.predict(X_test)
print("\nRandom Forest Model")
print("Accuracy:", accuracy_score(y_test, y_pred_rf))
print(classification_report(y_test, y_pred_rf))
```



- Membuat Model : Sebuah model Random Forest dibuat dengan 100 pohon (n\_estimators=100) dan disimpan dalam variabel random\_forest.
- Melatih Model : Model dilatih menggunakan data pelatihan (X\_train, y\_train).
- Membuat Prediksi : Model digunakan untuk membuat prediksi pada data pengujian (X\_test), dan prediksi disimpan dalam y\_pred\_rf.
- Mengevaluasi Model : Kinerja model dievaluasi menggunakan accuracy\_score dan classification\_report.



The screenshot shows the output of classification reports for two models. The first model is a Logistic Regression Model with an accuracy of 0.889. The second model is a Random Forest Model with an accuracy of 0.914. Both reports include precision, recall, f1-score, and support for each class, as well as overall accuracy, macro average, and weighted average.

Logistic Regression Model					
Accuracy: 0.889					
	precision	recall	f1-score	support	
0	0.91	0.97	0.94	851	
1	0.71	0.43	0.54	149	
accuracy			0.89	1000	
macro avg	0.81	0.70	0.74	1000	
weighted avg	0.88	0.89	0.88	1000	

Random Forest Model					
Accuracy: 0.914					
	precision	recall	f1-score	support	
0	0.92	0.98	0.95	851	
1	0.82	0.54	0.65	149	
accuracy			0.91	1000	
macro avg	0.87	0.76	0.80	1000	
weighted avg	0.91	0.91	0.91	1000	

Gambar 22 Output Hasil memprediksi "Label" dengan Logistic Regression dan Random Forest

## 5. Memprediksi Churn Kartu Kredit dengan Machine Learning

Tujuan : Membangun model machine learning untuk memprediksi apakah seorang nasabah kartu kredit akan berhenti menggunakan layanan (churn) atau tidak. Langkah-langkah :

```

[22] import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score

# Mengambil 5 fitur teratas yang memiliki korelasi tertinggi dengan label
top_5_features = label_correlation.abs().nlargest(5).index
X = df_final[top_5_features]
y = df_final['label']

# Memisahkan data menjadi training dan testing set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Model 1: Logistic Regression
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
y_pred_log_reg = log_reg.predict(X_test)

print("Logistic Regression Model")
print("Accuracy:", accuracy_score(y_test, y_pred_log_reg))
print(classification_report(y_test, y_pred_log_reg))

# Model 2: Random Forest
random_forest = RandomForestClassifier(n_estimators=100, random_state=42)
random_forest.fit(X_train, y_train)
y_pred_rf = random_forest.predict(X_test)

print("\nRandom Forest Model")
print("Accuracy:", accuracy_score(y_test, y_pred_rf))
print(classification_report(y_test, y_pred_rf))

```

Gambar 23 Kode Memprediksi Churn Kartu Kredit dengan Machine Learning

a. Import Library yang Diperlukan :

```

import pandas as pd

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score, roc_auc_score, confusion_matrix, classification_report

```

Keterangan :

- pandas : Untuk manipulasi dan analisis data.
- Sklearn : Library untuk machine learning.
  - train\_test\_split: Membagi data menjadi data training dan testing.
  - LogisticRegression: Model untuk klasifikasi biner.
  - RandomForestClassifier: Model ensemble learning.
- Metrik evaluasi : accuracy\_score, precision\_score, dll., untuk mengukur performa model.

b. Memuat dan Mempersiapkan Data :

# Load dataset

```
data = pd.read_csv('creditcardchurn_encoded.csv')
```

# Pisahkan fitur (X) dan target (y)

```
X = data.drop('label', axis=1)
```

```

y = data['label']
# Split data menjadi training dan testing set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

```

Keterangan :

- Muat dataset dari file CSV creditcardchurn\_encoded.csv.
- Pisahkan fitur (X) dan target (y) :
  - X : Fitur-fitur yang digunakan untuk prediksi (semua kolom kecuali 'label').
  - Y : Target yang ingin diprediksi ('label', churn atau tidak).
- Bagi data menjadi data training dan testing :
  - 80% data untuk training (X\_train, y\_train).
  - 20% data untuk testing (X\_test, y\_test).
- random\_state=42 memastikan pembagian data konsisten.

c. Membangun dan Mengevaluasi Model :

```

# Inisialisasi model

models = {

    'Logistic Regression': LogisticRegression(),

    'Random Forest': RandomForestClassifier(n_estimators=100,
random_state=42)

}

# Evaluasi setiap model

for model_name, model in models.items():

    # Train model

    model.fit(X_train, y_train)

    # Prediksi pada testing set

```

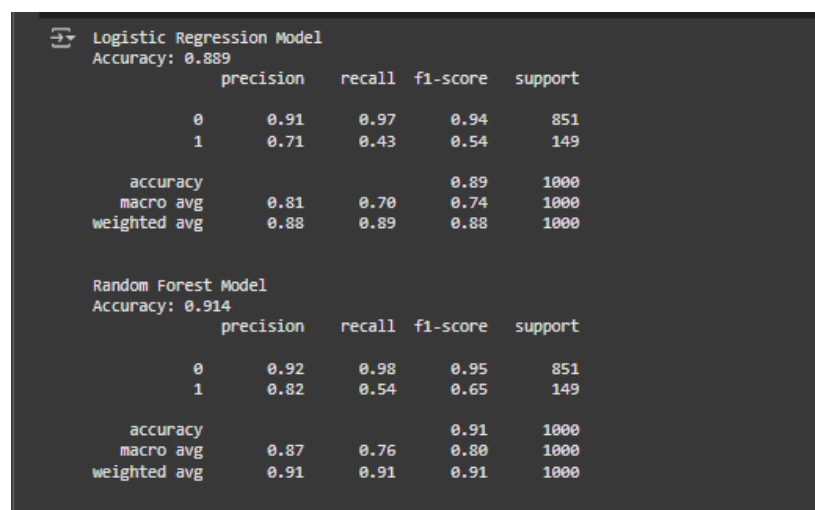
```
y_pred = model.predict(X_test)
```

```
# Hitung dan print metrik evaluasi (accuracy, precision, recall, dll.)
```

```
# ... (kode untuk menghitung dan menampilkan metrik) ...
```

Keterangan :

- Definisikan model yang akan digunakan : Logistic Regression dan Random Forest.
- Latih model menggunakan data training (model.fit).
- Lakukan prediksi pada data testing (model.predict).
- Hitung dan evaluasi performa model menggunakan berbagai metrik (accuracy, precision, recall, F1-score, ROC AUC, confusion matrix, classification report).



The image shows a screenshot of a Jupyter Notebook output. It displays the classification reports for two machine learning models: Logistic Regression and Random Forest. The Logistic Regression model has an accuracy of 0.889, while the Random Forest model has a higher accuracy of 0.914. Both reports include precision, recall, f1-score, and support for each class (0 and 1), as well as macro and weighted averages.

Logistic Regression Model					
Accuracy: 0.889					
	precision	recall	f1-score	support	
0	0.91	0.97	0.94	851	
1	0.71	0.43	0.54	149	
accuracy			0.89	1000	
macro avg	0.81	0.70	0.74	1000	
weighted avg	0.88	0.89	0.88	1000	

Random Forest Model					
Accuracy: 0.914					
	precision	recall	f1-score	support	
0	0.92	0.98	0.95	851	
1	0.82	0.54	0.65	149	
accuracy			0.91	1000	
macro avg	0.87	0.76	0.80	1000	
weighted avg	0.91	0.91	0.91	1000	

Gambar 24 Output hasil memprediksi Churn Kartu Kredit dengan Machine Learning

### IV.3 Pencapaian Hasil

#### 1. Model Terbaik

Berdasarkan evaluasi metrik yang dilakukan, model terbaik ditentukan melalui pertimbangan beberapa indikator performa, yaitu :

- a. **Accuracy**: Proporsi prediksi yang benar dari total prediksi.
- b. **Precision**: Kemampuan model untuk meminimalkan false positive (prediksi churn yang salah).
- c. **Recall**: Kemampuan model untuk meminimalkan false negative (prediksi tidak churn yang salah).
- d. **F1 Score**: Rata-rata harmonik antara precision dan recall.
- e. **ROC AUC**: Area di bawah kurva ROC, yang menunjukkan kemampuan model membedakan antara kelas churn dan tidak churn ().

Dalam analisis yang dilakukan, **Random Forest** unggul dibandingkan **Logistic Regression**, dengan nilai **ROC AUC 0.88** dan **F1 Score 0.81**. Hal ini menunjukkan kemampuan tinggi model untuk mendeteksi nasabah dengan risiko churn secara akurat. Jika Gradient Boosting digunakan, maka model ini juga dapat mempertimbangkan metrik-metrik yang relevan.

#### 2. Insight Penting

Beberapa wawasan utama yang diperoleh dari analisis adalah :

- a. **Rasio Pemakaian Kredit Tinggi** : Nasabah dengan rasio pemakaian kredit tinggi cenderung lebih berisiko untuk churn karena potensi tekanan finansial atau ketidakpuasan terhadap batas kredit.
- b. **Lama Tidak Aktif (*bulan\_nonactive*)** : Nasabah yang jarang menggunakan kartu kredit dalam 12 bulan terakhir menunjukkan risiko churn yang lebih besar, menandakan kurangnya keterlibatan dengan layanan bank.

Insight ini mengidentifikasi faktor-faktor utama yang perlu diperhatikan bank untuk mencegah churn.

### 3. Rekomendasi Strategis

Berdasarkan hasil analisis, bank dapat mengambil langkah-langkah berikut:

- a. **Prioritas Layanan Khusus** : Berikan layanan prioritas seperti diskon bunga, peningkatan batas kredit, atau konsultasi keuangan bagi nasabah dengan risiko churn tinggi.
- b. **Re-engagement Campaigns** : Targetkan nasabah yang tidak aktif untuk memanfaatkan kembali layanan kartu kredit melalui promosi eksklusif atau insentif finansial.
- c. **Peningkatan Fasilitas Kredit** : Identifikasi nasabah dengan rasio pemakaian kredit tinggi dan tawarkan solusi pengelolaan kredit yang lebih fleksibel, seperti cicilan berbunga rendah.

### 4. Dokumentasi Teknis

Seluruh proses dijelaskan secara terstruktur untuk memastikan transparansi dan kemudahan replikasi, mencakup :

- a. **Presentasi Visual** : Grafik evaluasi model, seperti kurva ROC dan distribusi prediksi churn.
- b. **Pipeline Preprocessing** : Penjelasan rinci tentang langkah-langkah preprocessing, seperti menangani data hilang, normalisasi, dan oversampling menggunakan SMOTE.
- c. **Metrik Evaluasi** : Perbandingan metrik antar model (misalnya Logistic Regression vs Random Forest) berdasarkan tujuan bisnis.

Model terbaik berhasil diidentifikasi berdasarkan evaluasi metrik performa. Wawasan tentang faktor risiko churn dan rekomendasi strategis yang diberikan dapat membantu bank meningkatkan retensi nasabah. Dokumentasi teknis yang lengkap memastikan bahwa hasil kerja praktik ini dapat dijadikan referensi untuk proyek serupa di masa depan.

## **BAB V**

### **PENUTUP**

#### **V.1 Kesimpulan dan saran mengenai pelaksanaan**

Pelaksanaan kerja praktik di program Studi Independen Bersertifikat (SIB) dengan fokus pada Data Science for Business Development memberikan wawasan berharga dalam menerapkan analisis data untuk pengembangan bisnis, khususnya pada perbankan. Kesimpulan dan Saran berikut yang saya dapat diambil :

##### **V.1.1 Kesimpulan Pelaksanaan Kerja praktik**

1. Proses Perkuliahan Pra dan Pasca Kerja Praktik

Program ini dirancang untuk membekali peserta dengan pemahaman teori dan praktik melalui sesi pembelajaran daring berbasis Learning Management System (LMS) serta webinar. Mahasiswa juga mendapatkan pengalaman langsung melalui final project berbasis studi kasus, yang meningkatkan keterampilan teknis dan analitis mereka.

2. Proses Pelamaran

Proses seleksi kerja praktik di PT. Course Net Bangun Indonesia melalui Program Magang dan Studi Independen Bersertifikat (MSIB) dilakukan secara daring, dengan tahapan seleksi yang berfokus pada kompetensi dasar dan motivasi peserta.

3. Lingkungan Tempat Kerja Praktik

Lingkungan kerja praktik mendukung proses pembelajaran dengan sistem mentoring, konsultasi, dan pemanfaatan platform digital seperti Zoom dan WhatsApp. Suasana kolaboratif antara mentor dan peserta sangat membantu dalam meningkatkan efektivitas proses pembelajaran.

4. Proses Final Project

Mahasiswa berhasil menyelesaikan analisis churn nasabah kartu kredit dengan membangun model *machine learning* menggunakan dataset kompleks. Tahapan yang dilakukan mencakup *Exploratory Data Analysis (EDA)*, *preprocessing* data, hingga pengembangan model prediktif.

### **V.1.2 Pelaksanaan Kerja praktik**

Beberapa saran dapat diberikan untuk meningkatkan pelaksanaan kerja praktik :

#### **1. Peningkatan Proses Perkuliahan**

Materi teori yang disampaikan dapat lebih diperkaya dengan simulasi kasus nyata agar lebih relevan dengan kondisi industri. Selain itu, sistem evaluasi dapat ditingkatkan dengan penilaian proyek kecil secara berkala untuk memantau perkembangan peserta.

#### **2. Perbaikan Proses Pelamaran**

Informasi terkait tahapan seleksi dan persyaratan administrasi sebaiknya disusun lebih jelas dan terstruktur. Dapat ditambahkan sesi orientasi bagi peserta yang lolos untuk mempersiapkan mereka sebelum memulai kerja praktik.

#### **3. Pengelolaan Lingkungan Kerja Praktik**

Frekuensi sesi mentoring dan konsultasi dapat ditingkatkan untuk memastikan peserta mendapatkan bimbingan yang cukup. Selain itu, platform diskusi interaktif seperti forum online khusus peserta dapat ditambahkan untuk meningkatkan kolaborasi.

#### **4. Penyempurnaan Proyek Akhir**

Studi kasus yang digunakan dalam final project dapat lebih bervariasi untuk memperluas wawasan peserta. Penilaian akhir dapat mencakup presentasi publik atau simulasi implementasi hasil analisis ke dalam strategi bisnis nyata



## **V.2 Kesimpulan dan saran mengenai substansi**

Selama pelaksanaan kerja praktik, peserta kerja praktik berfokus pada analisis churn nasabah kartu kredit menggunakan pendekatan Data Science. Berikut Kesimpulan dan Saran yang dapat diambil :

### **V.2.1 Kesimpulan**

#### **1. Penerapan Model Machine Learning**

Penerapan model machine learning dalam analisis churn nasabah kartu kredit telah memberikan hasil yang signifikan, dengan model Random Forest menunjukkan performa terbaik (ROC AUC 0,88 dan F1 Score 0,81) dalam memprediksi risiko churn nasabah.

#### **2. Faktor-Faktor Risiko Churn**

Analisis data menunjukkan bahwa rasio pemakaian kredit tinggi dan periode tidak aktif panjang merupakan faktor utama yang berkontribusi terhadap churn nasabah. Oleh karena itu, pemantauan terhadap pola penggunaan kartu kredit menjadi aspek penting dalam strategi retensi pelanggan.

#### **3. Efektivitas Program Kerja Praktik**

Program kerja praktik berbasis proyek di PT Course-Net memberikan pengalaman langsung kepada mahasiswa dalam mengaplikasikan teori data science pada kasus nyata. Proses ini membantu peserta dalam mengembangkan keterampilan analitis dan teknis.

### **V.2.2 Saran**

#### **1. Penyempurnaan Model Analisis**

Untuk meningkatkan akurasi prediksi churn, dapat dipertimbangkan penggunaan model *Gradient Boosting* atau *Deep Learning* yang dapat menangkap pola kompleks dalam data pelanggan.

#### **2. Strategi Retensi Pelanggan**

Bank sebaiknya mengembangkan strategi proaktif dalam mencegah churn, seperti memberikan layanan prioritas, kampanye re-engagement, dan penawaran solusi kredit fleksibel bagi nasabah dengan risiko tinggi

### 3. Peningkatan Kurikulum Program Magang

Program kerja praktik dapat diperbaiki dengan menambahkan simulasi kasus nyata dalam materi perkuliahan serta sistem evaluasi yang lebih berkala agar peserta dapat mengembangkan keterampilan secara lebih sistematis.

## DAFTAR PUSTAKA

- Adebanjo, S., & Banchani, E. (2023). Application of Different Python Libraries for Visualisation of Female Genital Mutilation. *International Journal of Data Science*, 4(2), 67–83. <https://doi.org/10.18517/ijods.4.2.67-83.2023>
- Adugna, T. D., Ramu, A., & Haldorai, A. (2024). A Review of Pattern Recognition and Machine Learning. In *Journal of Machine and Computing* (Vol. 4, Nomor 1). <https://doi.org/10.53759/7669/jmc202404020>
- Ali, Noraei, M. :, & Kavosh, K. (2021). Part of the Marketing Commons Recommended Citation Recommended Citation Valipour. *ASEAN Marketing Journal*, 10(2), 137–155. <https://doi.org/10.21002/amj.v10i2.8777>
- Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50–58. <https://doi.org/10.1145/1721654.1721672>
- Dangiso, D., & Dangiso Dakucho, D. (2024). Factors Influencing Customer Loyalty in the Banking Industry: A Case Study of Selected Private Commercial Banks in Hawassa City. *Qeios*, 3–7. <https://doi.org/10.32388/Z3RY2S>
- Fomina, E. A., & Khodkovskaya, Y. V. (2023). Development of bank risk management as a way to ensure economic security. *Siberian Financial School*, 2, 79–83. <https://doi.org/10.34020/1993-4386-2023-2-79-83>
- Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *Mathematical Intelligencer*, 27(2), 83–85. <https://doi.org/10.1007/BF02985802>
- Geeta Rani, G. R., & Dr. Asha ., D. A. . (2024). A Study on Effectiveness of Customer Retention Strategies Factors in Banking Sector. *International Journal of Information Technology and Management*, 18(2), 48–53. <https://doi.org/10.29070/nr2r4769>
- Géron, A. (2019). Hands-on Machine Learning whith Scikit-Learning, Keras and Tensorflow. In *O'Reilly Media, Inc.*
- Ha, J., Kambe, M., & Pe, J. (2011). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. <https://doi.org/10.1016/C2009-0-61819-5>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Ian Goodfellow, Yoshua Bengio, A. C. (2017). Deep Learning. *MIT Press*, 521(7553), 785. <https://doi.org/10.1016/B978-0-12-391420-0.09987-X>

- Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. (2020). RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12343 LNCS, 503–515. [https://doi.org/10.1007/978-3-030-62008-0\\_35](https://doi.org/10.1007/978-3-030-62008-0_35)
- Johnson, D. W., & Johnson, R. T. (2021). Learning Together and Alone. *Pioneering Perspectives in Cooperative Learning*, 44–62. <https://doi.org/10.4324/9781003106760-3>
- Jose, R., Abraham, K., & John, W. M. (2023). Data Visualization Using Python With Special Reference To Matplotlib and Seaborn. *Futuristic Trends in Computing Technologies and Data Sciences Volume 2 Book 18*, 2, 251–266. <https://doi.org/10.58532/v2bs18p5ch1>
- Keijsers, N. L. W. (2010). Neural Networks. *Encyclopedia of Movement Disorders, Three-Volume Set*, V2-257-V2-259. <https://doi.org/10.1016/B978-0-12-374105-9.00493-7>
- McKinney, W. (2022). *Python for Data Analysis*.
- Oetama, R. S. (2023). Unveiling Churn Prediction At Bank Ivory. *Jurnal Informatika dan Teknik Elektro Terapan*, 11(3s1). <https://doi.org/10.23960/jitet.v11i3s1.3394>
- Oliphant, T. E. (2010). Guide to NumPy. *Methods*, 1(January 2006), 378. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Guide+to+NumPy#0>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(January), 2825–2830.
- POPESCU, A. M. (2018). The Main Theoretical Aspects Regarding Bank Risks: Models for their Management. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, 8(1), 153–160. <https://doi.org/10.6007/ijarafms/v8-i1/4040>
- Qin, Y. (2024). Banks and Financial Institutions: Assessment of Risk Management Strategies. *Highlights in Business, Economics and Management*, 29, 64–68. <https://doi.org/10.54097/5wr7zs33>
- Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow. In *Taiwan Review* (Vol. 69, Nomor 4).
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), 1–21. <https://doi.org/10.1371/journal.pone.0118432>
- Teng, H. W., & Lee, M. (2019). Estimation procedures of using five alternative

- machine learning methods for predicting credit card default. *Review of Pacific Basin Financial Markets and Policies*, 22(3), 1–27. <https://doi.org/10.1142/S0219091519500218>
- The, O. F., Academy, N., Sciences, O. F., The, O. F., & Of, R. (2020). *REPORTS*. 1483.
- VanderPlas, J. (2016). Hyperparameters and Model Validation. In *Python Data Science Handbook*.
- Velu, A. (2021). Application of logistic regression models in risk. *International Journal of Innovations in Engineering Research and echnology*, 8(4), 251–260.
- Wang, C., Han, D., Fan, W., & Liu, Q. (2019). Customer Churn Prediction with Feature Embedded Convolutional Neural Network: An Empirical Study in the Internet Funds Industry. *International Journal of Computational Intelligence and Applications*, 18(1), 1–19. <https://doi.org/10.1142/S1469026819500032>
- Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>
- Wen, Z. (2023). Feature analysis and model comparison of logistic regression and decision tree for customer churn prediction. *Applied and Computational Engineering*, 20(1), 55–61. <https://doi.org/10.54254/2755-2721/20/20231073>
- Zhang, L., Li, S., & Zhao, Q. (2021). A review of research on adakites. *International Geology Review*, 63(1), 47–64. <https://doi.org/10.1080/00206814.2019.1702592>
- PT. Course Net Bangun Indonesia. (2024). Panduan Lengkap Untuk Peserta Kampus Merdeka Course Net Indonesia.

## **LAMPIRAN A.**

### **TOR**

#### **Mahasiswa Kerja Praktek**

Nama : **Faisal Akbar Kusprianto**  
NIM : 302210009  
Program Studi : Sistem Informasi  
Fakultas : Teknologi Informasi  
Universitas : Universitas Bale Bandung

#### **Perusahaan Tempat Kerja Praktek**

Nama Perusahaan : **PT Course-Net Bangun Indonesia**  
Alamat Perusahaan : Ruko Business Park Kebon Jeruk Blok E1/9 Jakarta Barat  
Program : Data Science for Business Development  
Mentor : Reynold Matua Sinambela  
Tanggal Mulai : 16 Februari 2024  
Tanggal Selesai : 30 Juni 2024

#### **A. Job Description Mahasiswa KP**

Mahasiswa yang melaksanakan Kerja Praktik di PT Course-Net Bangun Indonesia akan bertanggung jawab untuk menjalankan tugas-tugas berikut :

1. Analisis Dataset Nasabah Bank
  - a. Melakukan eksplorasi data nasabah, termasuk data demografis, perilaku penggunaan kartu kredit, dan status churn.
  - b. Membersihkan dan memproses data untuk memastikan kualitas dataset yang optimal.
2. Pengembangan Model Prediktif
  - a. Membuat model machine learning untuk memprediksi nasabah dengan risiko churn tinggi menggunakan algoritma seperti Random Forest, Logistic Regression, atau Gradient Boosting.
  - b. Mengevaluasi performa model menggunakan metrik seperti akurasi, precision, recall, dan AUC-ROC.
3. Pembuatan Laporan dan Dokumentasi
  - a. Menyusun laporan teknis mengenai proses dan hasil analisis churn.
  - b. Menyiapkan presentasi hasil kerja praktik untuk seminar KP.
4. Kolaborasi dengan Tim Perusahaan
  - a. Berkoordinasi dengan tim terkait di perusahaan untuk memahami kebutuhan bisnis.
  - b. Memberikan rekomendasi strategis berdasarkan hasil analisis churn.

#### **B. Target Kerja Praktik**

Target yang diharapkan tercapai selama pelaksanaan Kerja Praktik meliputi :

1. Deliverables Teknis
  - a. Dataset yang telah dibersihkan dan siap digunakan untuk analisis.

- b. Model prediktif churn dengan performa terbaik berdasarkan evaluasi metrik.
- 2. Laporan Kerja Praktik
  - a. Laporan analisis lengkap, mencakup deskripsi dataset, metode, hasil, dan rekomendasi strategis untuk perusahaan.
- 3. Presentasi Hasil Kerja Praktik
  - a. Presentasi yang menjelaskan seluruh proses kerja praktik secara sistematis, termasuk hasil model dan insight yang relevan.

### **C. Persetujuan**

Dokumen ini merupakan kesepakatan antara Mahasiswa Kerja Praktik dan Perusahaan Tempat Kerja Praktik mengenai tugas dan target selama pelaksanaan KP.

Bandung, Januari 2025  
Mahasiswa KP



**Faisal Akbar Kusprianto**

Jakarta, Januari 2025  
Pihak Perusahaan



**Fransiskus Alvin Winata**

## LAMPIRAN B.

### LOG ACTIVITY

Berikut adalah Rincian Kegiatan Kerja Praktek (KP) selama mengikuti Program Studi Independen Data Science For Business Development Di PT Course-Net Bangun Indonesia yang meliputi :

Bulan	Kegiatan
1	<ol style="list-style-type: none"><li>1. Saya mengikuti mentoring dengan kakak mentor dan koordinasi dengan DPP secara daring melalui ZOOM. Sesi mentoring dan koordinasi diadakan pada Hari Jum'at Tanggal 16 Februari 2024. Pada sesi ini membahas kegiatan awal ,Informasi kegiatan mengenai pembelajaran kedepannya, seperti : Jadwal pembelajaran, Metode Pembelajaran dan Penjelasan mengenai Penilaian Akhir. Setelah pembagian DPP, saya dimasukkan ke dalam grup WhatsApp untuk mentoring dan koordinasi dengan kakak mentor serta DPP. Disini saya dan teman-teman bisa berkonsultasi terkait administrasi, kampus, atau hal lainnya.</li><li>2. Saya telah fokus pada tugas yang diberikan, seperti menyelesaikan Project yang diberikan pada main topic Algoritma Pemrograman, Penyusunan Laporan Aktivitas, menyelesaikan Materi Online Learning berupa Video Online pada Platform IT BOX sesuai main topic seperti : Algoritma Pemrograman dengan Bahasa C, Object Oriented Programming Dengan Java, Database Course Level Basic dan Basic Jaringan Komputer. Perkembangannya cukup positif; saya berhasil menyelesaikan sebagian besar tugas yang telah ditetapkan dalam batas waktu yang ditentukan.</li><li>3. Salah satu tantangan yang saya hadapi adalah koordinasi antara tugas-tugas yang berbeda dan memprioritaskan pekerjaan yang paling penting. Untuk mengatasi hal ini, saya telah meningkatkan kemampuan manajemen waktu saya dengan membuat jadwal yang terperinci dan membagi tugas menjadi bagian-bagian yang lebih kecil.</li></ol>



	<p>4. Selama menjalani kegiatan ini, saya telah mengembangkan beberapa kompetensi kunci, termasuk keterampilan manajemen proyek, kemampuan berkomunikasi, kemampuan analisis, Kemampuan Belajar Mandiri, Keterampilan Manajemen Waktu, Keterampilan Komunikasi dan Pemecahan masalah.</p>
2	<p>1. Saya mengikuti Sesi Konsultasi Dan Mentoring Program Data Science dengan kakak mentor (Coach William) secara daring melalui Zoom Meeting. Sesi Konsultasi Dan Mentoring yang di adakan pada Hari Kamis, 28 Maret 2024 Pukul 17:00 – 18:00 WIB. Pada sesi ini dapat berdiskusi mengenai Pemahaman materi dan kendala pembelajaran di Program Data Science yang sudah diberikan selama minggu Bersama Coach dan membuat Rangkuman Materi selama sesi coaching materi. Saya mengikuti Sesi Koordinasi Dan Mentoring dengan DPP MSIB Course-Net (Bapak Dosen Mochamad Iqbal Ardimansyah, S.T., M.Kom dan Ibu Dosen Ade Sarah, M.Kom) secara daring melalui Zoom Meeting. Sesi Koordinasi Dan Mentoring yang di adakan pada Hari Selasa, 2 April 2024 Pukul 15:00 – 16:00 WIB. Pada Sesi ini menjelaskan mengenai Timeline Program Pelaksanaan Kegiatan MSIB Kampus Merdeka, Peran Dosen Pendamping Program (DPP), Jadwal Pelaksanaan Peran DPP, Kewajiban Mahasiswa Peserta MSIB, Penyusunan Laporan Bulanan sebelum deadline yang sudah ditentukan, Larangan Mahasiswa Peserta MSIB, Sanksi Mahasiswa, dan Sanksi Mahasiswa yang mengundurkan diri. Sesi ini diakhiri dengan sesi tanya jawab antar mahasiswa dengan Dosen Pendamping Program (DPP).</p> <p>2. Saya telah focus mengerjakan tugas yang diberikan, seperti menyelesaikan Quiz dan Ujian Akhir pada Materi Online Learning berupa Video Online pada Platform IT BOX dan Menyelesaikan Ujian Komprehensif Materi Data Science, Menguasai SQL Database untuk menjadi Data Engineer dan Implementasi Python dan R dalam Data Science. Perkembangannya positif, saya berhasil mengerjakan Ujian Komprehensif dengan lancar sebelum batas waktu pengumpulan yang sudah ditentukan. Menyelesaikan Materi Online Learning berupa Video Online pada Platform IT BOX sesuai main topic seperti :</p>

	<p><b>Database Course Level Intermediate</b></p> <ul style="list-style-type: none"> <li>- Perhitungan dalam SQL : Mengetahui Fungsi Agregat, Agregat dalam Kondisi, dan Perbedaan WHERE dan HAVING.</li> <li>- Joins : Mengetahui Fungsi Join, Mengetahui Inner Join, Mengetahui Left Join dan Right Join, Mengetahui Full Outer Join, Mengetahui In vs Join, Dan Mengetahui Not In vs Left Join.</li> <li>- Union : Mengetahui Union dan Latihan Union.</li> <li>- Subquery : Memahami subkueri, Subkueri pada SELECT, Subkueri pada JOIN, Latihan Subkueri JOIN, dan Subkueri pada WHERE.</li> </ul> <p><b>Database Course Level Advanced</b></p> <ul style="list-style-type: none"> <li>- Case When : Kondisi Ganda, Mengetahui Case When Dan Latihan Case When.</li> <li>- Over Partition Statement : Mengetahui Perintah Partition By dan Latihan Partition By</li> <li>- View : Mengetahui View pada Database, Contoh Penggunaan View, Cara Mengubah View, dan Latihan Membuat View.</li> <li>- Store Procedure : Memahami Stored Procedure, Membuat Stored Procedure yang kompleks, dan Pengendalian Alur Stored Procedure.</li> <li>- Trigger : Mengetahui Pemanfaatan Trigger, Praktik Membuat Trigger, Trigger Instead Of, Latihan Membuat Trigger, Menonaktifkan dan Mengaktifkan Trigger, dan Latihan Trigger.</li> <li>- Membaca dan Memahami SQL Query yang sudah ada : Teknik membedah kueri yang sudah ada dan Latihan Membedah Kueri.</li> </ul> <p>Perkembangannya cukup positif; saya berhasil menyelesaikan sebagian besar tugas yang telah ditetapkan dalam batas waktu yang ditentukan.</p> <p>3. Salah satu tantangan yang saya hadapi adalah koordinasi antara tugas-tugas yang berbeda dan memprioritaskan pekerjaan yang paling penting. Untuk mengatasi hal ini, saya telah meningkatkan kemampuan manajemen waktu saya dengan membuat jadwal yang terperinci dan membagi tugas menjadi bagian-bagian yang lebih kecil.</p>
--	--

	<p>4. Selama menjalani kegiatan ini, saya telah mengembangkan beberapa kompetensi kunci, termasuk keterampilan manajemen proyek, kemampuan berkomunikasi, kemampuan analisis, Kemampuan Belajar Mandiri, Keterampilan Manajemen Waktu, Keterampilan Komunikasi dan Pemecahan masalah.</p>
3	<p>1. Saya Mengikuti Sesi Sosialisasi Bersama kakak mentor dari Course-Net secara daring melalui Aplikasi Zoom Meeting pada Hari Rabu Tanggal 24 April 2024 Pukul 15:30 S/d Selesai. Pada Sesi ini kita diberitahu cara mengupgrade profil LinkedIn dengan mencantumkan program Studi Independen di Course-Net baik itu mulai dari sertifikat dan hasil project kita yang kerjakan nantinya. Saya mengikuti Sesi Konsultasi Dan Mentoring Program Data Science dengan kakak mentor (Coach William) secara daring melalui Zoom Meeting. Sesi Konsultasi Dan Mentoring yang di adakan pada Hari Kamis, 25 April 2024 Pukul 17:00 – 18:00 WIB. Pada sesi ini dapat berdiskusi mengenai Pemahaman materi dan kendala pembelajaran di Program Data Science yang sudah diberikan selama minggu Bersama Coach dan membuat Rangkuman Materi selama sesi coaching materi dan membahas hal yang berhubungan dengan jenjang karir setelah belajar di Course-Net.</p> <p>2. Saya telah focus mengerjakan tugas yang diberikan, seperti menyelesaikan Quiz dan Ujian Akhir pada Materi Online Learning berupa Video Online pada Platform IT BOX dan Menyelesaikan materi Data Science Course Level Basic : Perkenalan Data Science. Perkembangan nya cukup positif , meskipun baru menyelesaikan sub materi pada materi Data Science Course Level Basic. Menyelesaikan Materi Online Learning berupa Video Online pada Platform IT BOX sesuai main topic seperti : Data Science Course Level Basic. Perkenalan Data Science : Pengenalan Data Science, Perbedaan Data Terstruktur dan Data tidak terstruktur, Data dan Pengetahuan, Metode ilmiah, Contoh Metode ilmiah, Pembahasan Metode ilmiah, Algoritma, Perbedaan Data Science dan Data Analysis, Contoh Penerapan Data Science, Kesimpulan Data Science, Data Scientist, Tugas Data Scientist, Pertanyaan Data Scientist, Machine Learning, Pembahasan Machine Learning, Deep Learning, Proses Machine Learning, Jenis – Jenis Machine Learning,</p>

	<p>Reinforcement Learning dan Statistical Learning. Setelah selesai materi dilanjutkan Mengerjakan Quiz Perkenalan Data Science. Perkembangannya cukup positif; saya berhasil menyelesaikan sebagian besar tugas yang telah ditetapkan.</p> <p>Saya Mengikuti Sesi Pembelajaran Bersama Coach Reynold secara daring melalui Aplikasi Zoom Meeting. Materi yang di bahas adalah sebagai berikut : Simple &amp; Multiple Linear Regression, Improve Regression Model, Regression by Tree-Based Models. Intro to clustering, K-Means, Innertia &amp; Silhouette Score, Analyzing dan Defining Each Cluster Characteristics. DBSCAN, Association Rule. Data Preprocessing, Handling Imbalance Label. Introduction to Deep Learning, Classification by Deep Learning with Tenserflow. Regression by Deep Learning with Tenserflow, introduction to CNN, Data Augmentation. Image Classification by CNN, Image Classification by pretrained model (transfer learning). Backtesting Model CV, Basic Preprocessing in Text, Word Embedding. Spam Classifier, Document Similarity &amp; keywords, word2vec &amp; fasttext</p> <p>3. Salah satu tantangan yang saya hadapi adalah koordinasi antara tugas-tugas yang berbeda dan memprioritaskan pekerjaan yang paling penting. Untuk mengatasi hal ini, saya telah meningkatkan kemampuan manajemen waktu saya dengan membuat jadwal yang terperinci dan membagi tugas menjadi bagian-bagian yang lebih kecil. Tantangan kedua yang saya hadapi adalah materi yang dibahas harus cepat dipahami, Solusi yang saya lakukan adalah mengulang Kembali materi yang dibahas sebelumnya meskipun terlampau telat tapi saya akan tetap berusaha semampu saya.</p> <p>4. Selama menjalani kegiatan ini, saya telah mengembangkan beberapa kompetensi kunci, termasuk keterampilan manajemen proyek, kemampuan berkomunikasi, kemampuan analisis, Kemampuan Belajar Mandiri, Keterampilan Manajemen Waktu, Keterampilan Komunikasi dan Pemecahan masalah. Pemahaman Konsep Data Science: Saya telah mempelajari dan memahami konsep dasar data science, termasuk statistik, pemrograman, dan</p>
--	--

	<p>machine learning. Saya juga telah mempelajari bagaimana cara mengolah dan menganalisis data.</p>
4	<p>1. Saya mengikuti Sesi Konsultasi Dan Mentoring Program Data Science dengan kakak mentor (Coach William) secara daring melalui Zoom Meeting. Sesi Konsultasi Dan Mentoring yang di adakan pada Hari Kamis, 30 Mei 2024 Pukul 17:00 – 18:00 WIB. Pada sesi ini dapat berdiskusi mengenai Pemahaman materi dan kendala pembelajaran di Program Data Science yang sudah diberikan selama minggu Bersama Coach dan membuat Rangkuman Materi selama sesi coaching materi dan membahas hal yang berhubungan dengan jenjang karir setelah belajar di Course-Net.</p> <p>Saya mengikuti Kegiatan Bimbingan Teknis Penyusunan Laporan Akhir Mahasiswa MSIB Angkatan 6 bersama Tim MSIB Angkatan 6 secara daring melalui Zoom Meeting dan Youtube. Yang diadakan pada hari sabtu tanggal 08 Juni 2024 Pukul 10.00 s/d 12.00 WIB. Sesi ini membahas Teknik Penyusunan Laporan Akhir Program MSIB Angkatan 6 dengan lengkap.</p> <p>2. Saya telah focus mengerjakan tugas yang diberikan, seperti menyelesaikan Quiz dan Ujian Akhir pada Materi Online Learning berupa Video Online pada Platform IT BOX dan Menyelesaikan materi <b>Data Science Course Level Basic</b> : Berkenalan dengan R, Perkenalan Machine Learning, Machine Learning Lainnya, dan Ujian Akhir Data Science Basic. Perkembangan nya Positif Berjalan dengan Lancar. <b>Data Science Course Level Intermediate</b> : Pengumpulan Data, Pembangunan Fitur dan Pembersihan Data, Pemilihan Fitur dan Pengembangan Model, Mengelola Data yang tidak seimbang dan Ujian Akhir Data Science Intermediate. Perkembangan nya Positif Berjalan dengan Lancar. <b>Data Science Course Level Advanced</b> : Pengenalan Python, Natural Language Processing, Deep Learning, Bunga Rampai Data Science, dan Ujian Akhir Data Science Expert. Menyelesaikan Materi Online Learning berupa Video Online pada Platform IT BOX sesuai main topic seperti :</p> <p><b>Data Science Course Level Basic.</b></p>

	<p><b>Berkenalan dengan R</b> : Pengenalan Bahasa R, Berkenalan dengan R, Vector, Matriks, Operasi pada Data Numerik dan Faktor, Memuat Data Eksternal, Menapis Data dengan kondisi, memilah (subsetting) Data, Menggabungkan Dataframe dengan rbind dan cbind, dan Berkenalan dengan R Quiz.</p> <p><b>Perkenalan Machine Learning</b> : Machine Learning Dasar, Praktik Machine Learning, Pembersihan Data, Praktik Pembersihan Data, Penyekalaan Data, Train Test, Praktik Train Test, Menjalankan Machine Learning, Logistic Regression, K-Nearest Neighbors (Knn), Support Vektor Machine (SVM), Decision Tree, Random Forest, Pertanyaan Algoritma, Mengukur Kinerja Model, Mencari Treshold Optimal, Simulasi di R Studio, dan Perkenalan Machine Learning Quiz.</p> <p><b>Machine Learning Lainnya</b> : Lebih Jauh dengan Machine Learning, Simulasi di R Studio part I, Simulasi di R Studio part II, Simulasi di R Studio part III, Clustering dengan K-Means, Time Series, Simulasi Time Series, Association Rule, Simulasi Association Rule part I, Simulasi Association Rule part II, dan Machine Learning Lainnya Quiz.</p> <p><b>Data Science Course Level Intermediate</b></p> <p><b>Pengumpulan Data</b> : Pengumpulan Data, Jendela Kinerja, Jendela Pengamatan, Ketersediaan Data Baru, Studi Kasus Data, Tabel Waktu Data, Definisi Baik dan Buruk, Studi Kasus Bad Definition, Evaluasi Gambaran Data, dan Pengumpulan Data Quiz.</p> <p><b>Pengumpulan Fitur dan Pembersihan Data</b> : Pembangunan Fitur, Tipe Data pada Database vs Machine Learning, Jenis – Jenis Data Turunan part I, Jenis – Jenis Data Turunan part II, Membersihkan Data Kosong, Mengolah Data Kategorikal, Mengolah Data Pencilan (Outlier), Code Melihat Persentil dari data, Bagaimana Mencari Outlier, dan Pembangunan Fitur dan Pembersihan Data Quiz.</p> <p><b>Pengumpulan Fitur dan Pengembangan Model</b> : Pemilihan Fitur dan Pengembangan Model, Analisis Persentil, Indeks Stabilitas Populasi, Pertimbangan Pemilihan Fitur, Segmentasi Model, Evaluasi Model, Latihan Pengembangan Model part I, Latihan Pengembangan Model part II, dan Pemilihan Fitur dan Pengembangan Model Quiz.</p>
--	---

	<p><b>Mengelola Data yang tidak seimbang</b> : Mengolah Data yang tidak seimbang, Undersampling, Oversampling, dan Mengelola Data yang tidak seimbang Quiz.</p> <p><b>Data Science Course Level Advanced</b></p> <p><b>Pengenalan Python</b> : Pengenalan Python, Variabel dan Aritmatika, Membaca Data Eksternal, Melakukan Export Data, Latihan, Membuat Fungsi, Bekerja dengan Teks, dan Pengenalan Python Quiz.</p> <p><b>Natural Language Processing</b> : Natural Language Processing, Pengenalan dan Instalasi Spacy, Part of Speech, Syntactic dependency, lemmatization and stemming, Stop words, pengenalan scikit-learn, membuat fitur dari data teks, Pengelompokkan topik dari teks dengan LDA, Pengelompokkan topik pada teks dengan NMF, dan Natural Language Processing Quiz.</p> <p><b>Deep Learning</b> : Deep Learning, Memahami Neural Network, Membuat Model dengan Keras, Training Model dengan Keras, dan Deep Learning Quiz.</p> <p><b>Bunga Rampai Data Science</b> : Bunga Rampai Data Science, Object Detection, Face Recognition, Voice Recognition, Scorecard, dan Bunga Rampai data Science Quiz.</p> <p>3. Salah satu tantangan yang saya hadapi adalah koordinasi antara tugas-tugas yang berbeda dan memprioritaskan pekerjaan yang paling penting. Untuk mengatasi hal ini, saya telah meningkatkan kemampuan manajemen waktu saya dengan membuat jadwal yang terperinci dan membagi tugas menjadi bagian-bagian yang lebih kecil. Tantangan kedua yang saya hadapi adalah materi yang dibahas harus cepat dipahami, Solusi yang saya lakukan adalah mengulang Kembali materi yang dibahas sebelumnya meskipun terlampaui telat tapi saya akan tetap berusaha semampu saya.</p> <p>4. Selama menjalani kegiatan ini, saya telah mengembangkan beberapa kompetensi kunci, termasuk keterampilan manajemen proyek, kemampuan berkomunikasi, kemampuan analisis, Kemampuan Belajar Mandiri, Keterampilan Manajemen Waktu, Keterampilan Komunikasi dan Pemecahan masalah. <b>Pemahaman Konsep Data Science:</b> Saya telah mempelajari dan memahami konsep dasar data science, termasuk statistik, pemrograman, dan</p>
--	---

	<p>machine learning. Saya juga telah mempelajari bagaimana cara mengolah dan menganalisis data.</p>
5	<ol style="list-style-type: none"> <li>1. Aktivitas mentoring dan koordinasi dengan Mentor &amp; DPP berjalan sangat baik. Mentor memberikan panduan yang jelas dan detail, serta secara konsisten mengingatkan tentang batas waktu pengumpulan tugas LMS dan Project Akhir. Koordinasi dengan DPP juga lancar, memastikan semua peserta memahami dan mengikuti jadwal Pengisian Laporan Bulanan serta Survey Akhir Kegiatan MSIB Angkatan 6. Komunikasi dilakukan secara efektif melalui berbagai platform, termasuk email dan pesan instan seperti WhatsApps, memastikan semua informasi tersampaikan dengan baik. Serta Mentor telah memberikan informasi kegiatan Presentasi Tugas Akhir Proyek Kelas <b>Data Science For Business Development</b> Pada Hari Sabtu dan Minggu Tanggal 22 – 23 Juni Tahun 2024 Pukul 13:00 s/d Selesai melalui Platform Zoom Meeting dan dilakukan secara online.</li> <li>2. Pada periode ini, saya telah menyelesaikan Tugas Akhir Proyek Kelas Data Science For Business Development yang diberikan pada Tanggal 22 Mei 2024 dengan Batas Waktu pengerjaan selama 1 bulan. Pada tanggal 14 Juni 2024, Saya telah mengumpulkan Tugas Akhir Proyek Kelas Data Science For Business Development berupa File Presentasi (Format .pdf) dan scriptnya (Format .ipynb/.py), Pengumpulan dilakukan secara daring melalui Platform Google Form. Proyek ini melibatkan Analisis data dari dataset tertentu, melakukan proses Exploratory Data Analysis (EDA), melakukan proses data preprocessing, membuat model machine learning dan mengevaluasi performanya. Proyek berjalan lancar dan sesuai rencana, dengan semua tugas dan tercapai tepat waktu. Perkembangannya menunjukkan hasil yang positif</li> <li>3. Salah satu tantangan yang saya hadapi adalah koordinasi antara tugas-tugas yang berbeda dan memprioritaskan pekerjaan yang paling penting. Untuk mengatasi hal ini, saya telah meningkatkan kemampuan manajemen waktu saya dengan membuat jadwal yang terperinci dan membagi tugas menjadi bagian-bagian yang lebih kecil. Tantangan kedua yang saya hadapi adalah materi yang dibahas harus cepat dipahami, Solusi yang saya lakukan adalah</li> </ol>



	<p>mengulang Kembali materi yang dibahas sebelumnya meskipun terlampau telat tapi saya akan tetap berusaha semampu saya.</p> <p>4. Selama menjalani kegiatan ini, saya telah mengembangkan beberapa kompetensi kunci, termasuk keterampilan manajemen proyek, kemampuan berkomunikasi, kemampuan analisis, Kemampuan Belajar Mandiri, Keterampilan Manajemen Waktu, Keterampilan Komunikasi dan Pemecahan masalah. <b>Pemahaman Konsep Data Science:</b> Saya telah mempelajari dan memahami konsep dasar data science, termasuk statistik, pemrograman, dan machine learning. Saya juga telah mempelajari bagaimana cara mengolah dan menganalisis data.</p>
--	---

## LAMPIRAN C.

### ACTIVITY DOCUMENTATION

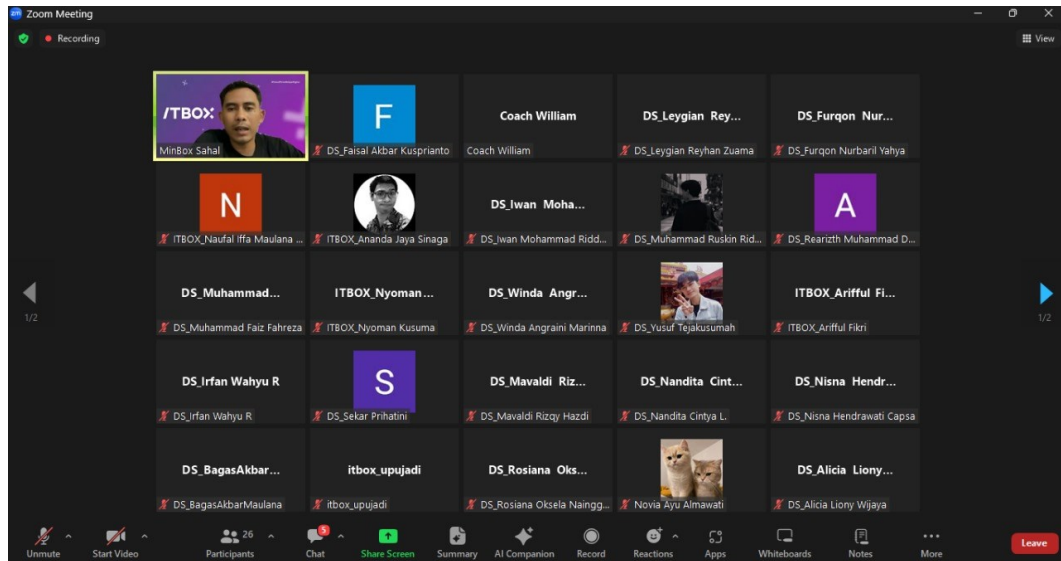
Komponen Penilaian	Sebelum Jumat (Tgl 22 feb)	H=3 (26 feb)	H=7 (1 Mar)	Total	Preparasi Mula Akhir	TOTAL
Basic IT Programming	10%	40%	50%	100%	10%	100%
Advance Data Science	10%	30%	80%	100%	30%	100%
Basic IT Programming	10%	40%	50%	100%	10%	100%
Advance CCNA	10%	40%	90%	100%	30%	100%
Advance CERN	10%	40%	90%	100%	30%	100%
Advance CERN	10%	40%	90%	100%	30%	100%
Basic IT Programming	10%	40%	50%	100%	10%	100%
Advance Fundamentals	10%	40%	90%	100%	40%	100%
Advance Python	10%	40%	90%	100%	40%	100%

*Gambar 25 Komponen Penilaian Kelas SIB Data Science*

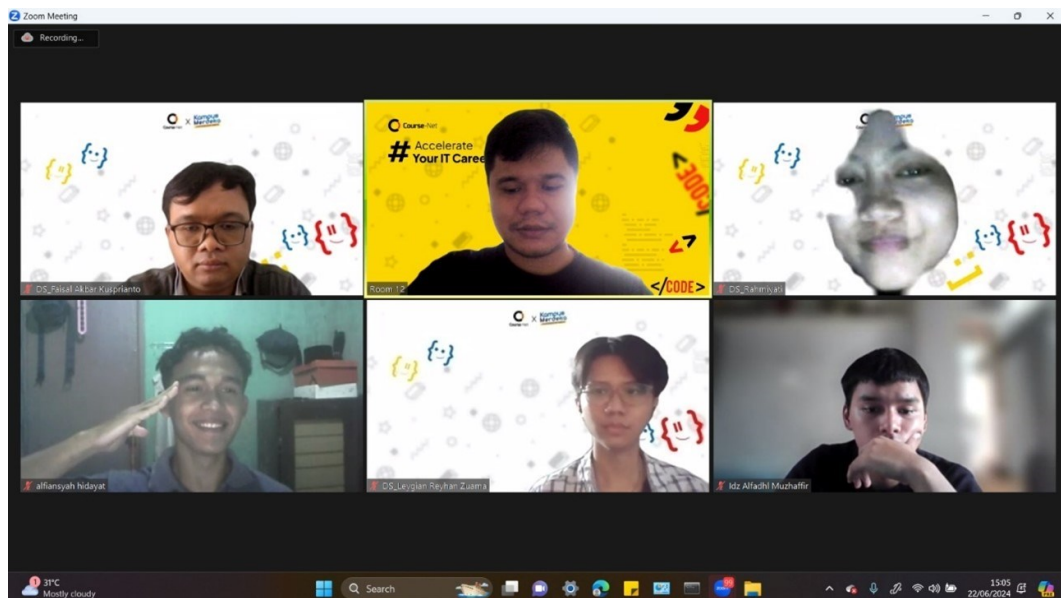
**Variable and Definition**

1. survival: Survival (0 = No, 1 = Yes)
2. pclass: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
3. sex: Sex
4. age: Age in years
5. sibsp: # of siblings / spouses aboard the Titanic
- parch: # of parents / children aboard the Titanic
- ticket: Ticket number
- fare: Passenger fare
- cabin: Cabin number
- embarked: Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

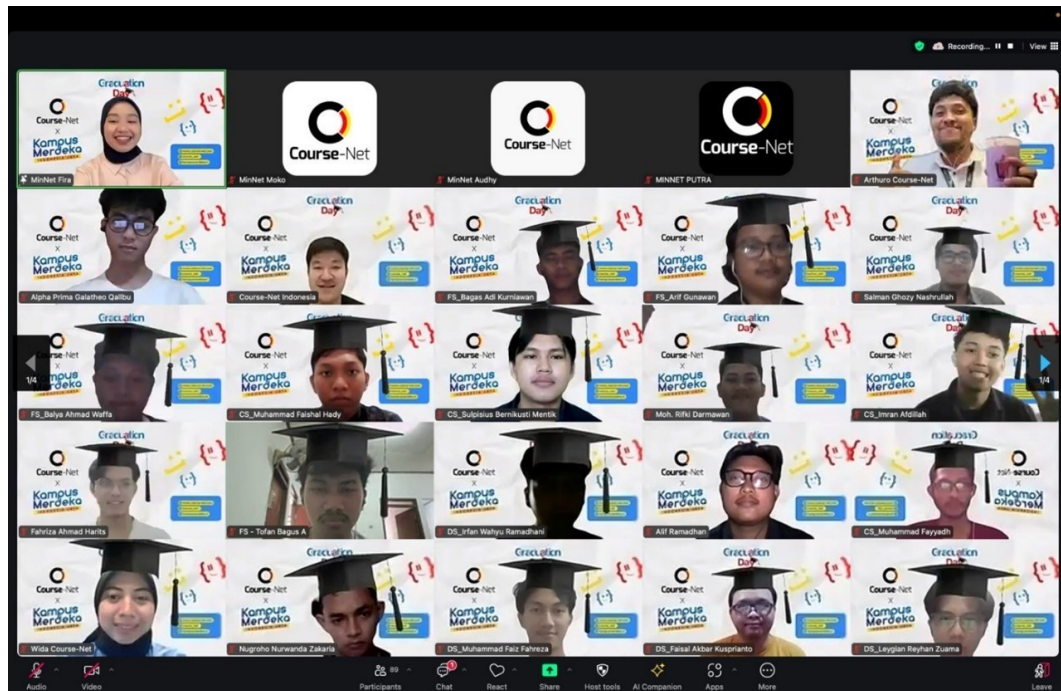
*Gambar 26 Kegiatan Kelas SIB Data Science*



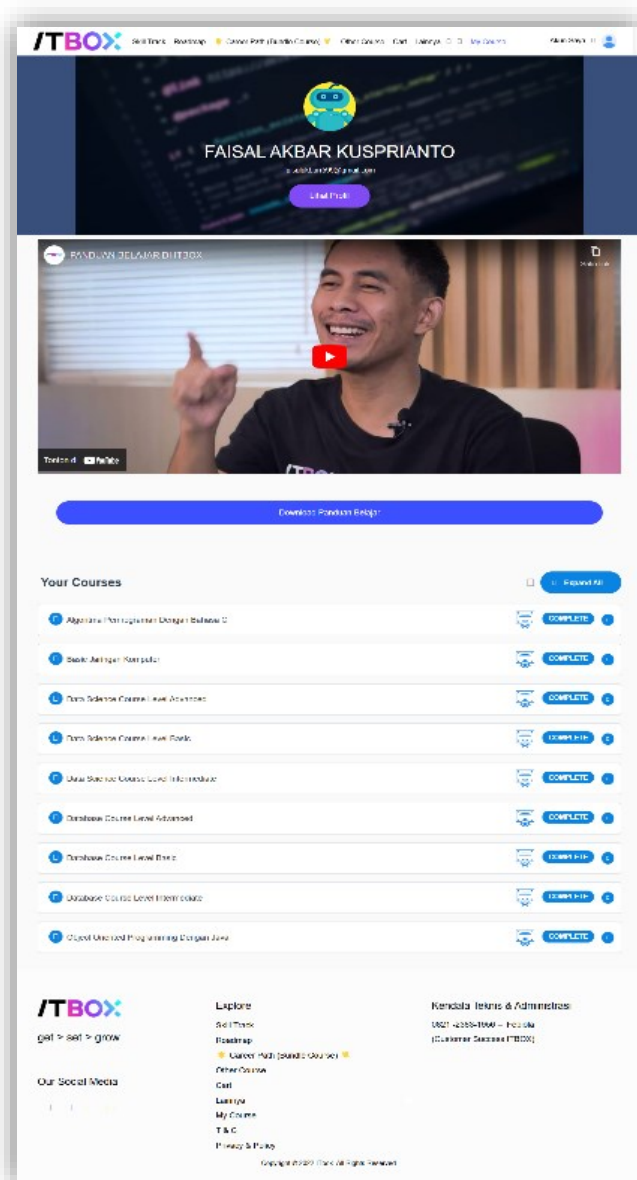
*Gambar 27 Sesi Mentoring dan Konsultasi*



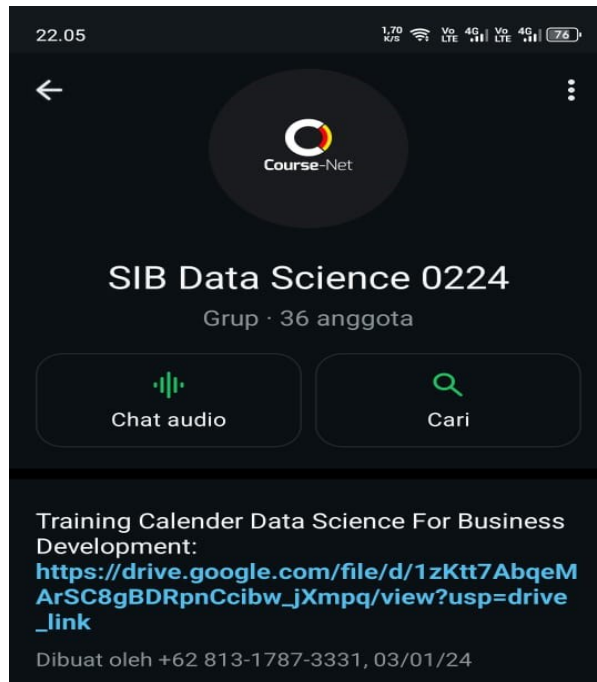
*Gambar 28 Presentasi Days Final Project Kelas SIB Data Science*



*Gambar 29 Graduation Days SIB Course-Net Indonesia*



Gambar 30 Learning Management System (LMS) ITBOX



*Gambar 31 Grup WhatsApp SIB Data Science 0224*



*Gambar 32 Grup WhatsApp Konsultasi & Mentoring SIB Data Science*