2024 - V1

GISELLE'S DATA ADVENTURE

DATA ANALYSIS AND MACHINE LEARNING

By Giselle Halim

HELLO! I'M
GISELLE.

Fueled by a passion for data and machine learning, I'm dedicated to harnessing the power of analytics to drive actionable insights. My journey has been enriched by experiences at Kalbis University and the prestigious Bangkit Academy, where I've honed my skills in data analysis, visualization, and machine learning. I've further solidified my expertise by earning the Google Data Analytics Certification.

Equipped with a solid foundation in data analysis tools like Excel, SQL, Power BI, and Tableau, I've refined my skills through practical projects. My technical expertise extends to Python, Scikit-learn, and TensorFlow, where I've explored the potential of machine learning algorithms. I'm eager to contribute my expertise to a dynamic organization as a data analyst or data scientist, where I can leverage data-driven insights to innovate and drive positive outcomes.

EDUCATION

Kalbis University (2020 - 2024)

Information Systems - Big Data Analytics

GPA: 3.94 / 4.00

Achievements:

- National Finalist of ASEAN Data Science Explorers 2024
- 2nd place - Internal competition for Information System Analysis and Design course

Organization:

Kalbis University Information Systems Student Association (HIMSI GALAKSI)

EXPERIENCES

Bangkit Academy 2023
Machine Learning Student

Graduated with a score of 93.5/100. For the capstone team project, **conceptualized and developed an Android-based educational application for traditional fabric motif identification using Machine Learning** techniques, reaching 93% accuracy in recognizing diverse motifs.

Orbit Future Academy
AI Mastery Student

Graduated with a score of 88/100. For the capstone team project, **spearheaded the development of an AI model for a web application designed to detect chili plant diseases**, successfully training the model on 500 images and achieving a detection accuracy rate of 82%.

EXPERIENCES

Data Scientist Virtual Intern

Nov - Dec 2023

Home Credit Indonesia x Rakamin Academy

Developed a credit risk prediction model using company-provided loan data using Random Forest and reached 90% accuracy.

Data Scientist Virtual Intern

Sep - Oct 2023

ID/X Partners x Rakamin Academy

Worked on building a credit risk prediction model for a lending company using loan data using Random Forest and reached 90% accuracy.

Data Scientist Virtual Intern

Aug - Sep 2023

Kalbe Nutritionals x Rakamin Academy

Enhanced business strategies by creating a Tableau dashboard and developing predictive models for customer segmentation using K-Means Clustering.

Big Data Analytics Virtual Intern

Jul - Aug 2023

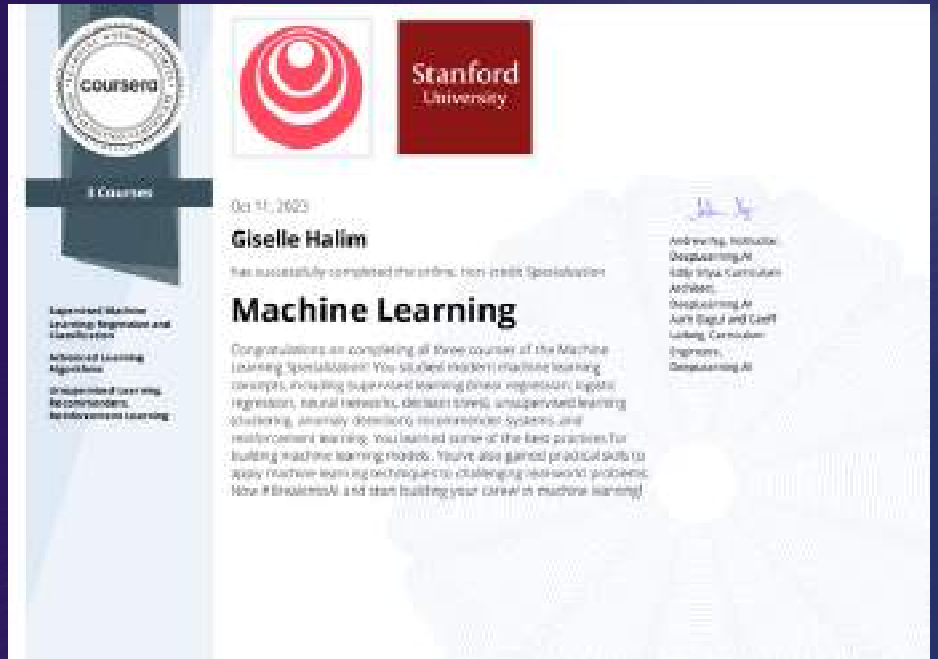
Kimia Farma x Rakamin Academy

Analyzed data using SQL to generate insights and visualized findings using Google Data Studio. Created a comprehensive dashboard to track medicine sales from raw data.

TECHNICAL SKILLS

- **Languages:** Python, SQL
- **Libraries:** Pandas, Numpy, Scikit-Learn, TensorFlow, Keras, Matplotlib, Seaborn
- **Web Technologies:** HTML, Flask, Streamlit
- **Tools:** Power BI, Tableau, Looker, SAP Analytics Cloud, MySQL, Microsoft Excel
- **Data Analysis:** Data Cleaning, Exploratory Data Analysis (EDA), Predictive Analysis, Cluster Analysis, Sentiment Analysis
- **Machine Learning:** Predictive Modeling, Image Classification, NLP, Recommender System

CERTIFICATES



coursera

3 Courses

Supervised Machine Learning: Regression and Classification

Advanced Learning Algorithms

Unsupervised Learning, Recommenders, Reinforcement Learning

Stanford University

Oct 11, 2023

Giselle Halim

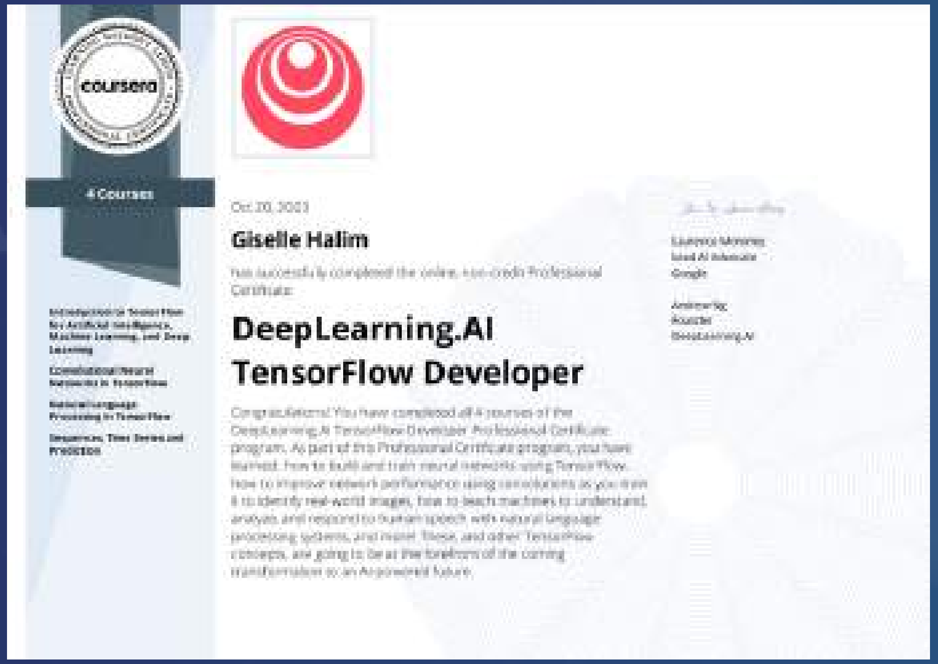
has successfully completed the online, non-credit Specialization

Machine Learning

Congratulations on completing all three courses of the Machine Learning Specialization! You studied modern machine learning concepts, including supervised learning (linear regression, logistic regression, neural networks, decision trees), unsupervised learning (clustering, anomaly detection), recommender systems, and reinforcement learning. You learned some of the best practices for building machine learning models. You've also gained practical skills to apply machine learning techniques to challenging real-world problems. Now #BreakIntoAI and start building your career in machine learning!

Andrew Ng, Instructor, DeepLearning.AI
Eddy Shyu, Curriculum Architect, DeepLearning.AI
Aarti Bagul and Geoff Ladwig, Curriculum Engineers, DeepLearning.AI

Machine Learning Specialization



coursera

4 Courses

Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning

Convolutional Neural Networks in TensorFlow

Natural Language Processing in TensorFlow

Sequences, Time Series and Prediction

Oct 20, 2023

Giselle Halim

has successfully completed the online, non-credit Professional Certificate

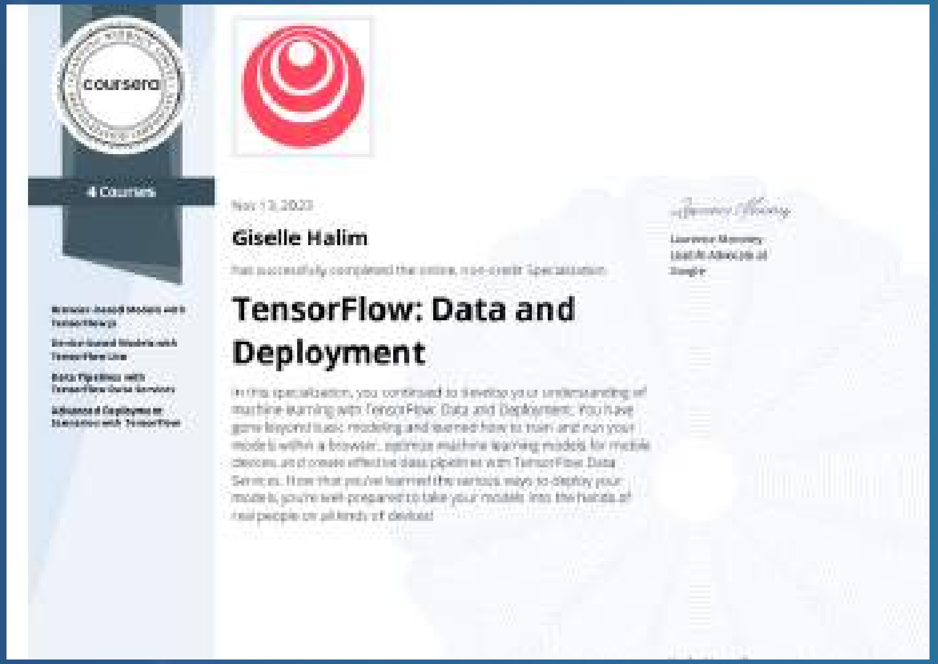
DeepLearning.AI TensorFlow Developer

Congratulations! You have completed all 4 courses of the DeepLearning.AI TensorFlow Developer Professional Certificate program. As part of this Professional Certificate program, you have learned: how to build and train neural networks using TensorFlow, how to improve network performance using convolutions as you train it to identify real-world images, how to teach machines to understand, analyze, and respond to human speech with natural language processing systems, and more! These, and other TensorFlow concepts, are going to be at the forefront of the coming transformation to an AI-powered future.

Laurence Moroney
Lead AI Advocate
Google

Andrew Ng
Founder
DeepLearning.AI

TensorFlow Developer by DeepLearning.AI



coursera

4 Courses

Browser-based Models with TensorFlow.js

Device-based Models with TensorFlow Lite

Data Pipelines with TensorFlow Data Services

Advanced Deployment Scenarios with TensorFlow

Nov 13, 2023

Giselle Halim

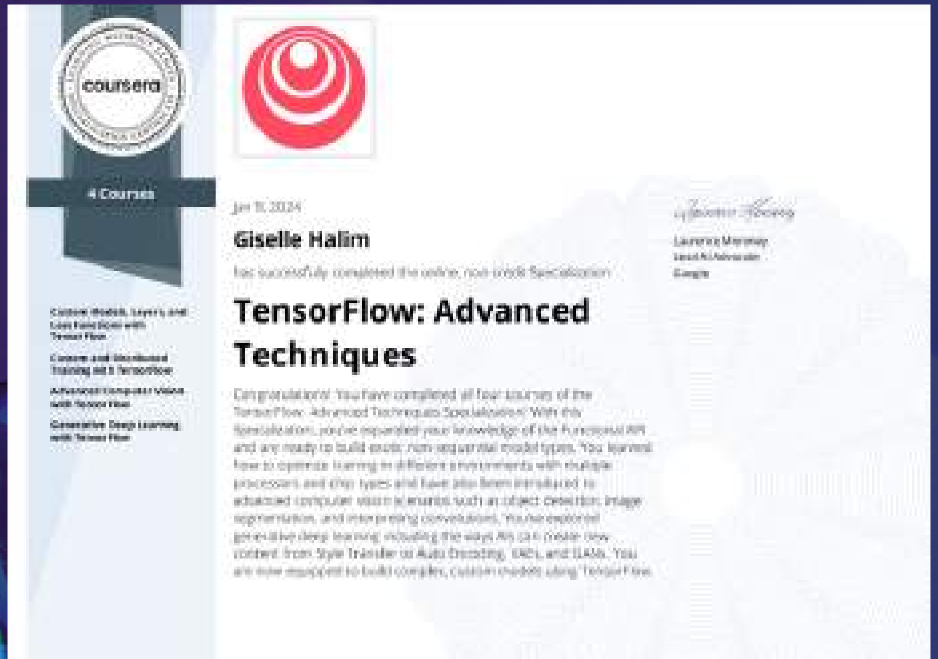
has successfully completed the online, non-credit Specialization

TensorFlow: Data and Deployment

In this specialization, you continued to develop your understanding of machine learning with TensorFlow: Data and Deployment. You have gone beyond basic modeling and learned how to train and run your models within a browser, optimize machine learning models for mobile devices, and create effective data pipelines with TensorFlow Data Services. Now that you've learned the various ways to deploy your models, you're well-prepared to take your models into the hands of real people on all kinds of devices!

Laurence Moroney
Lead AI Advocate at
Google

TensorFlow: Data and Deployment



coursera

4 Courses

Custom Models, Layers, and Loss Functions with TensorFlow

Custom and Distributed Training with TensorFlow

Advanced Computer Vision with TensorFlow

Generative Deep Learning with TensorFlow

Jan 11, 2024

Giselle Halim

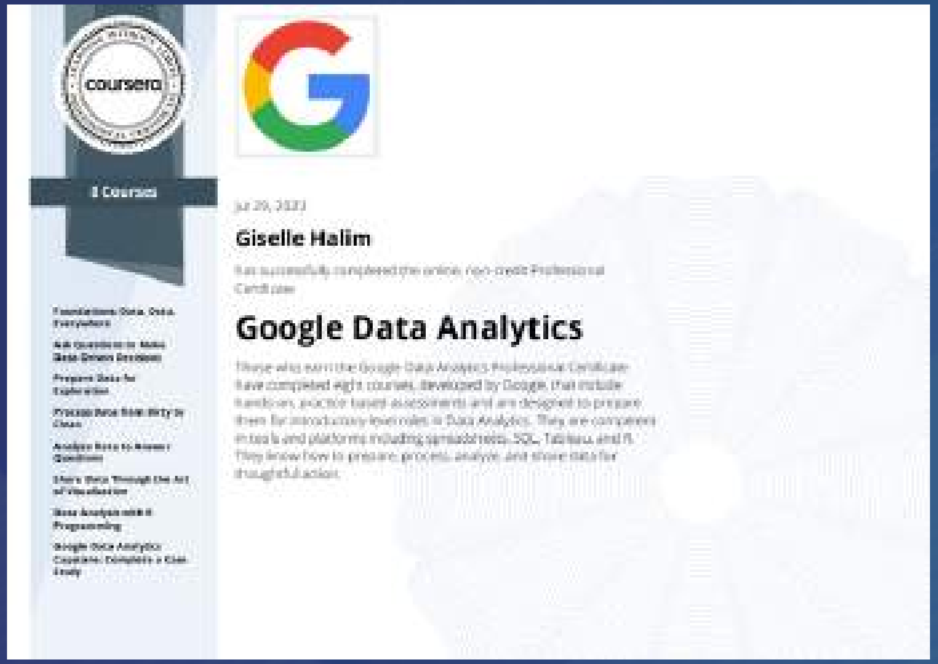
has successfully completed the online, non-credit Specialization

TensorFlow: Advanced Techniques

Congratulations! You have completed all four courses of the TensorFlow: Advanced Techniques Specialization! With this Specialization, you've expanded your knowledge of the Functional API and are ready to build exotic non-sequential model types. You learned how to optimize training in different environments with multiple processors and chip types and have also been introduced to advanced computer vision scenarios such as object detection, image segmentation, and interpreting convolutions. You've explored generative deep learning including the ways AIs can create new content from Style Transfer to Auto Encoding, VAEs, and GANs. You are now equipped to build complex, custom models using TensorFlow.

Laurence Moroney
Lead AI Advocate
Google

TensorFlow: Advanced Techniques



coursera

8 Courses

Foundations: Data, Data, Everywhere

Ask Questions to Make Data-Driven Decisions

Prepare Data for Exploration

Process Data from Dirty to Clean

Analyze Data to Answer Questions

Share Data Through the Art of Visualization

Data Analysis with R Programming

Google Data Analytics Capstone: Complete a Case Study

Jul 29, 2023

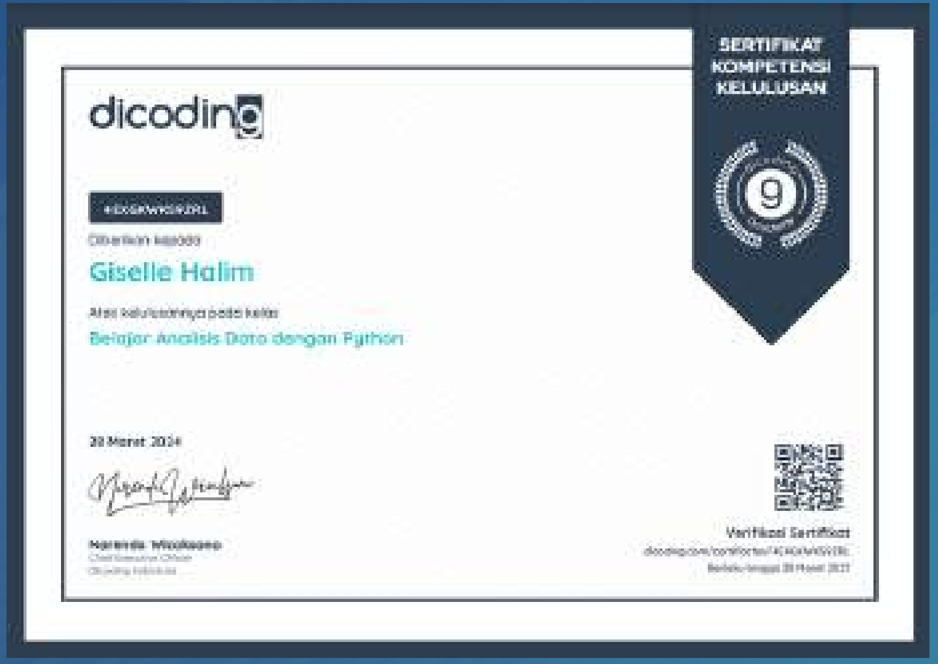
Giselle Halim

has successfully completed the online, non-credit Professional Certificate

Google Data Analytics

Those who earn the Google Data Analytics Professional Certificate have completed eight courses, developed by Google, that include hands-on, practice-based assessments and are designed to prepare them for introductory-level roles in Data Analytics. They are competent in tools and platforms including spreadsheets, SQL, Tableau, and R. They know how to prepare, process, analyze, and share data for thoughtful action.

Google Data Analytics Professional Certificate



dicoding

SERTIFIKAT KOMPETENSI KELULUSAN

[illegible]

Diberikan kepada

Giselle Halim

Atas kelulusannya pada kelas

Belajar Analisis Data dengan Python

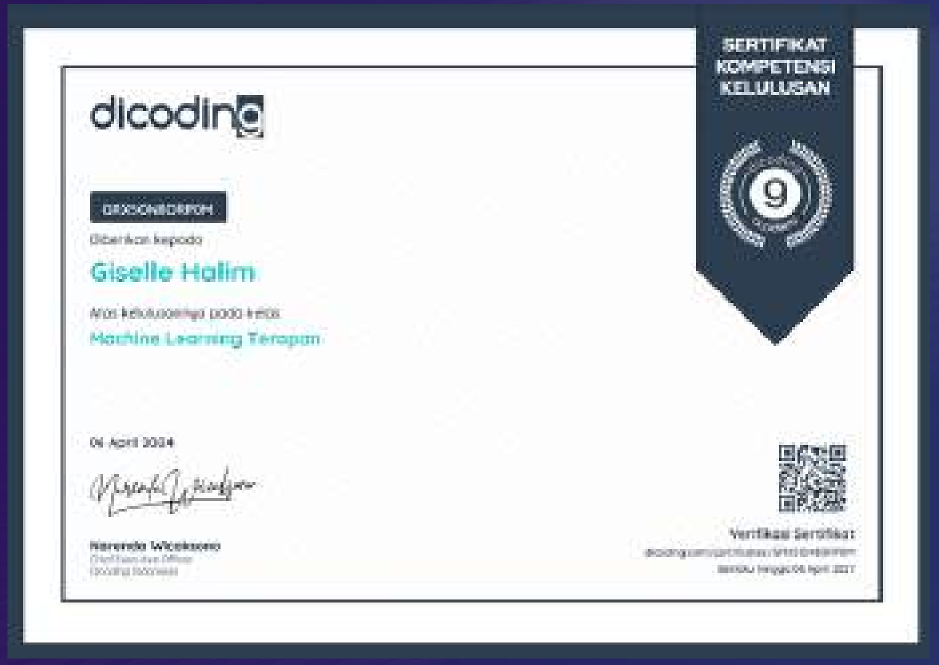
20 Maret 2024

Narenda Wicaksono
Chief Executive Officer
Dicoding Indonesia

Verifikasi Sertifikat
dicoding.com/certificates/[illegible]
Berlaku hingga 20 Maret 2027

Data Analysis with Python

CERTIFICATES



SERTIFIKAT KOMPETENSI KELULUSAN

dicoding

Diberikan kepada

Giselle Halim

Atas kelulusannya pada kelas

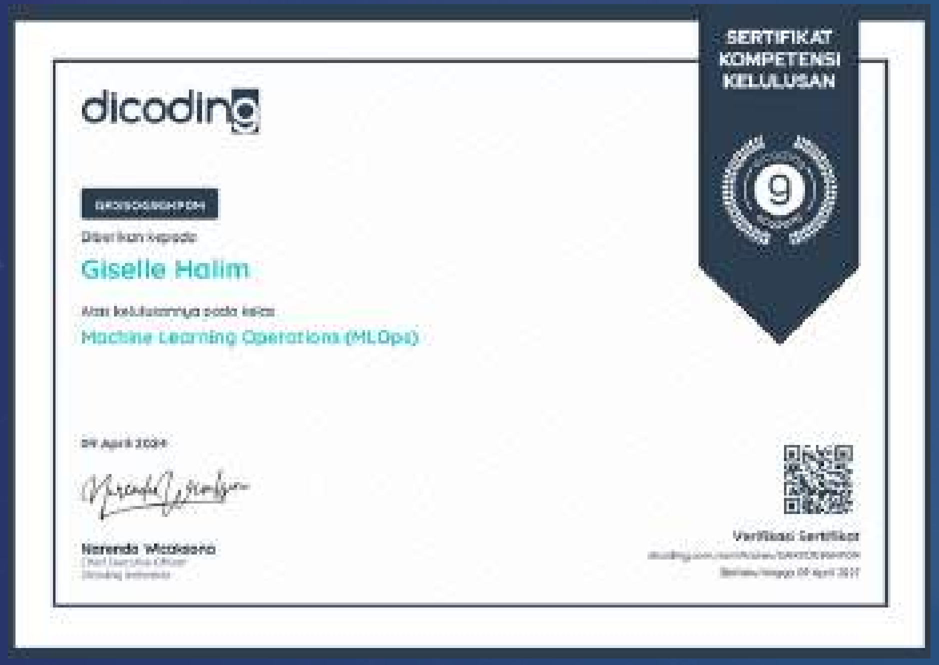
Machine Learning Terapan

06 April 2024

Narenda Wicaksono
Chief Executive Officer
Dicoding Indonesia

Verifikasi Sertifikat

Machine Learning Implementation



SERTIFIKAT KOMPETENSI KELULUSAN

dicoding

Diberikan kepada

Giselle Halim

Atas kelulusannya pada kelas

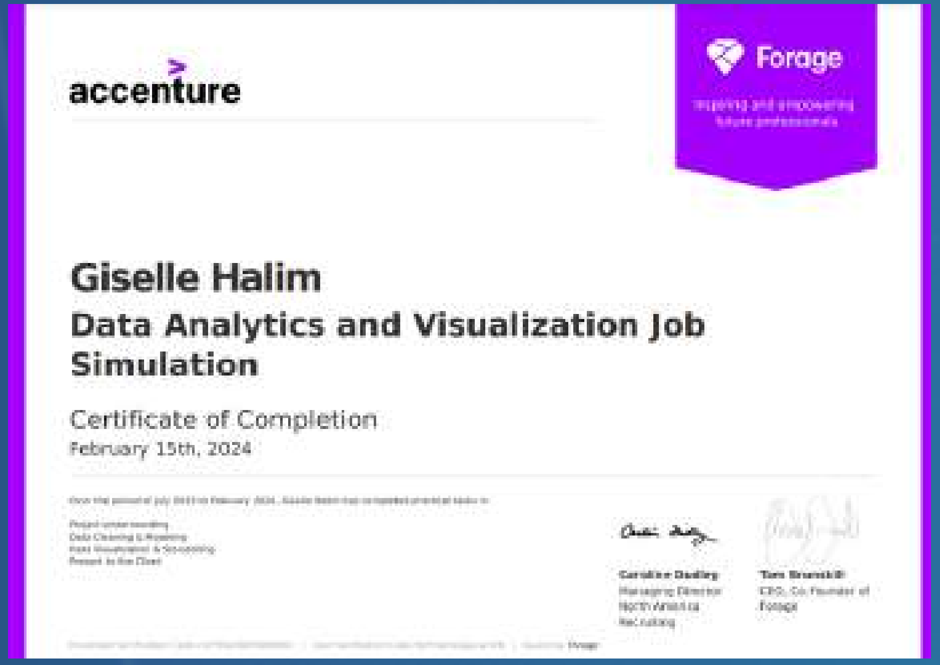
Machine Learning Operations (MLOps)

09 April 2024

Narenda Wicaksono
Chief Executive Officer
Dicoding Indonesia

Verifikasi Sertifikat

Machine Learning Operations (MLOps)



accenture

Forage
Inspiring and empowering future professionals

Giselle Halim

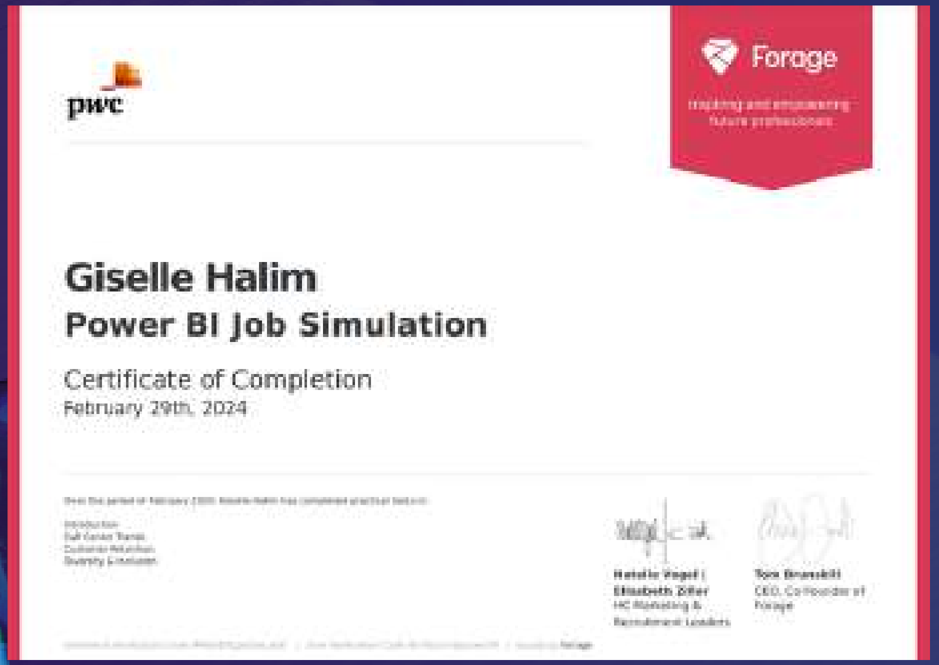
Data Analytics and Visualization Job Simulation

Certificate of Completion
February 15th, 2024

Caroline Dudley
Managing Director
North America
Recruiting

Tom Brunskill
CEO, Co-Founder of
Forage

Accenture Data Analytics Job Simulation



pwc

Forage
Inspiring and empowering future professionals

Giselle Halim

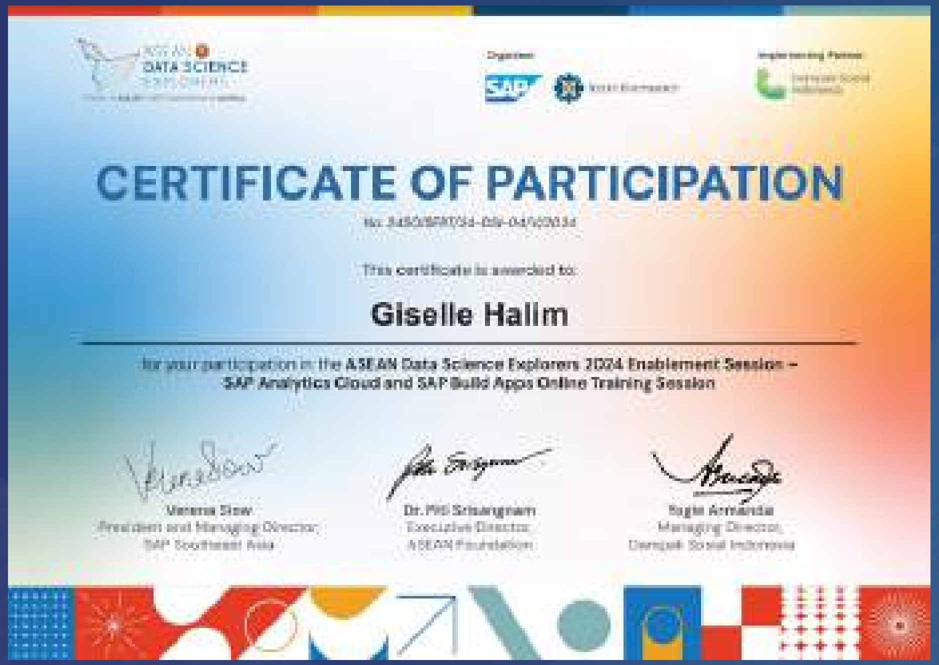
Power BI Job Simulation

Certificate of Completion
February 29th, 2024

Natalie Vogel |
Elisabeth Ziller
HC Marketing &
Recruitment Leaders

Tom Brunskill
CEO, Co-Founder of
Forage

PwC Switzerland Power BI Job Simulation



ASEAN DATA SCIENCE EXPLORERS

Organizer:

SAP ASEAN Foundation

Implementing Partner:

Dampak Sosial Indonesia

CERTIFICATE OF PARTICIPATION

This certificate is awarded to:

Giselle Halim

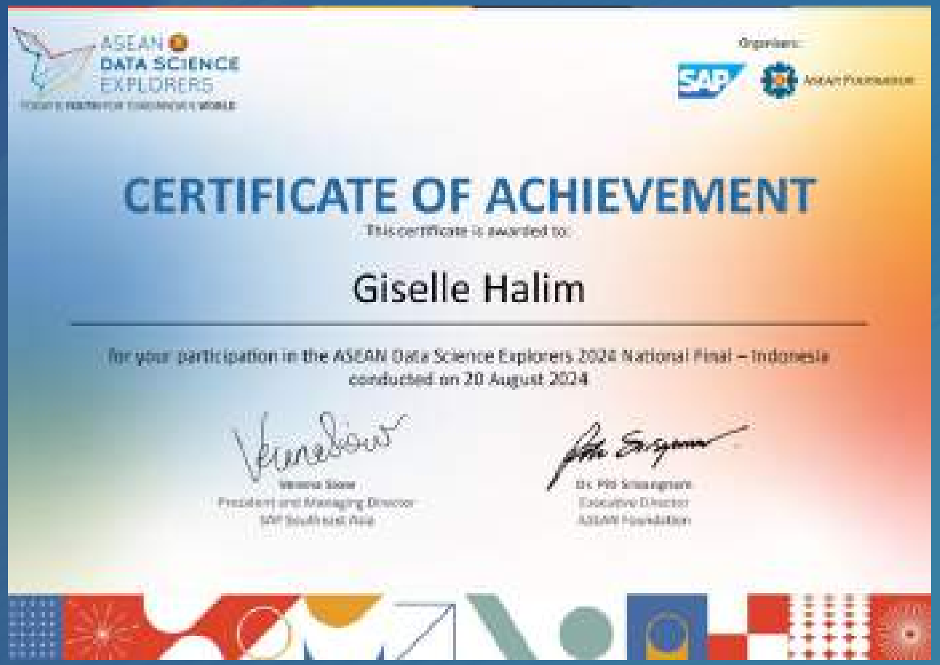
for your participation in the **ASEAN Data Science Explorers 2024 Enablement Session – SAP Analytics Cloud and SAP Build Apps Online Training Session**

Verena Siow
President and Managing Director
SAP Southeast Asia

Dr. Piti Srisangnam
Executive Director
ASEAN Foundation

Yogie Armanda
Managing Director
Dampak Sosial Indonesia

ASEAN DSE 2024 Enablement Session



ASEAN DATA SCIENCE EXPLORERS

TODAY'S YOUTH FOR TOMORROW'S WORLD

Organisers:

SAP ASEAN Foundation

CERTIFICATE OF ACHIEVEMENT

This certificate is awarded to:

Giselle Halim

for your participation in the ASEAN Data Science Explorers 2024 National Final – Indonesia conducted on 20 August 2024

Verena Siow
President and Managing Director
SAP Southeast Asia

Dr. Piti Srisangnam
Executive Director
ASEAN Foundation

ASEAN DSE 2024 National Finalist

PROJECTS

Showcasing past works related to **data analysis, data science, and machine learning**

DIABETES IN ASEAN ANALYSIS

DIABETES: SILENT ENEMY

-About diabetes, its risk factors, effects, and solutions. A disease that can hurt in silence, and do great damage when everyone looks away-

Presented By:
HUNTER
- Giselle Halim (Kalbis University)
- Michelle Angelini (Universitas Bunda Mulia)

03
Diabetes Prevalence per Country (2011 - 2021)
Undiagnosed Diabetes Prevalence 2021

Diabetes is a growing health crisis in ASEAN, with an average prevalence of 8.6%. Not only that, a significant portion of the population remains undiagnosed, particularly in countries like Indonesia where the rate reaches 73.7%.

The increasing number of diabetes cases can lead to more health problems, which can bring losses to a country if it occurs on a large scale. In addition, the high rate of undiagnosed diabetes can also be detrimental in the long run.

Source: IDF Diabetes Atlas

Source: IDF Diabetes Atlas

Deaths Related to Diabetes (2021)

Mortality Related to Diabetes

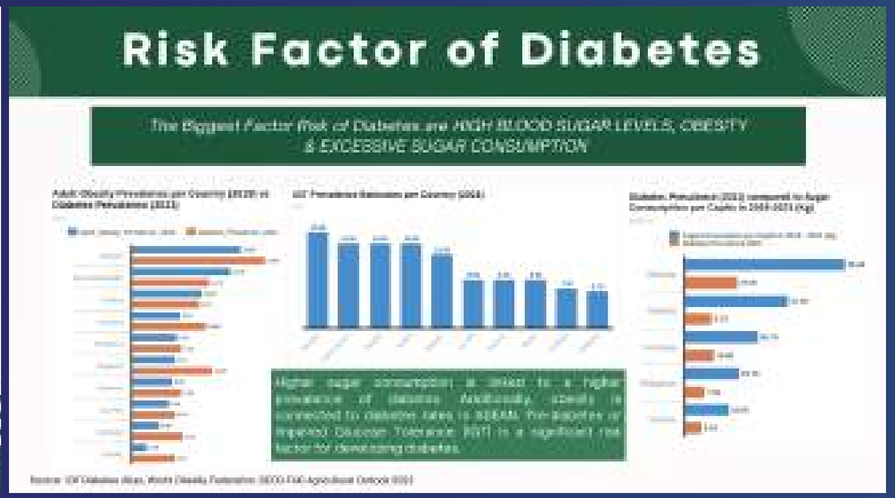
Diabetes-related mortality rates are on an upward trend worldwide, signaling the need for IMMEDIATE ACTION.

In addition to disease complications, **diabetes also takes a heavy toll on lives**. This is shown through the number of diabetes-related deaths. This number is increasing in all regions.

Within ASEAN countries, Indonesia had the highest rank of death that caused by Diabetes at 236,711 deaths.

Source: IDF Diabetes Atlas

Source: IDF Diabetes Atlas

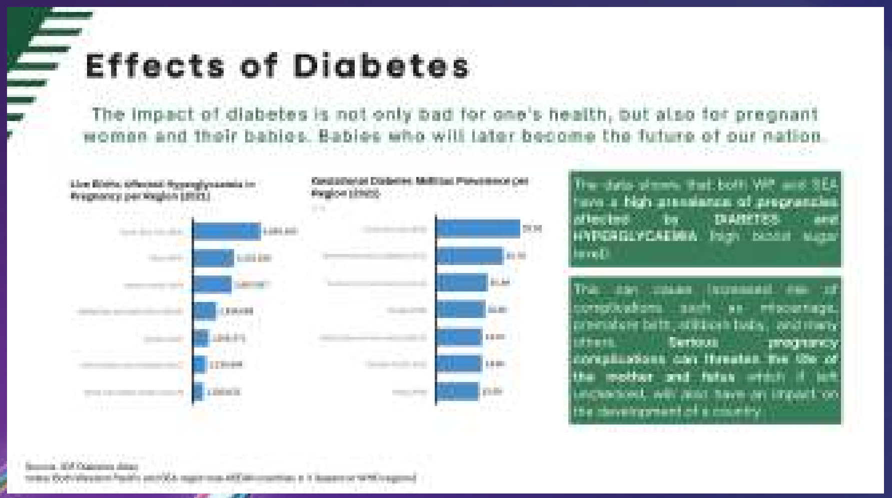

Adult Obesity Prevalence per Country (2020) vs Diabetes Prevalence (2021)
IGT Prevalence Estimates per Country (2021)
Diabetes Prevalence (2021) compared to Sugar Consumption per Capita in 2019-2021 (Kg)

Risk Factor of Diabetes

The Biggest Factor Risk of Diabetes are HIGH BLOOD SUGAR LEVELS, OBESITY & EXCESSIVE SUGAR CONSUMPTION

Higher sugar consumption is linked to a higher prevalence of diabetes. Additionally, obesity is connected to diabetes rates in ASEAN. The presence of Impaired Glucose Tolerance (IGT) is a significant risk factor for developing diabetes.

Source: IDF Diabetes Atlas, World Obesity Federation, OECD-FAO Agricultural Outlook 2023


Live Births Affected Hyperglycaemia in Pregnancy per Region (2021)
Gestational Diabetes Mellitus Prevalence per Region (2021)

Effects of Diabetes

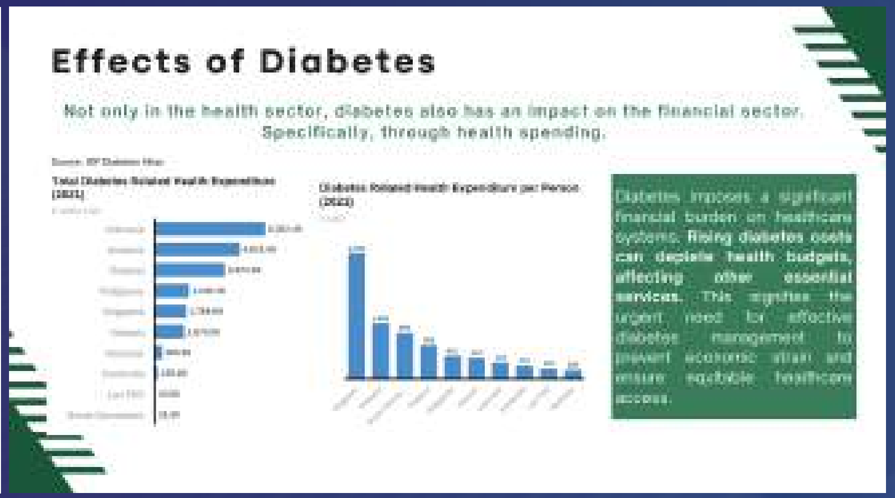
The impact of diabetes is not only bad for one's health, but also for pregnant women and their babies. Babies who will later become the future of our nation.

The data shows that both WP and SEA have a **high prevalence of pregnancies affected by DIABETES and HYPERGLYCAEMIA** (high blood sugar level).

This can cause increased risk of complications such as miscarriage, premature birth, stillborn baby, and many others. **Serious pregnancy complications can threaten the life of the mother and fetus** which if left unchecked, will also have an impact on the development of a country.

Source: IDF Diabetes Atlas

[illegible]


Total Diabetes Related Health Expenditure (2021)
Diabetes Related Health Expenditure per Person (2021)

Effects of Diabetes

Not only in the health sector, diabetes also has an impact on the financial sector. Specifically, through health spending.

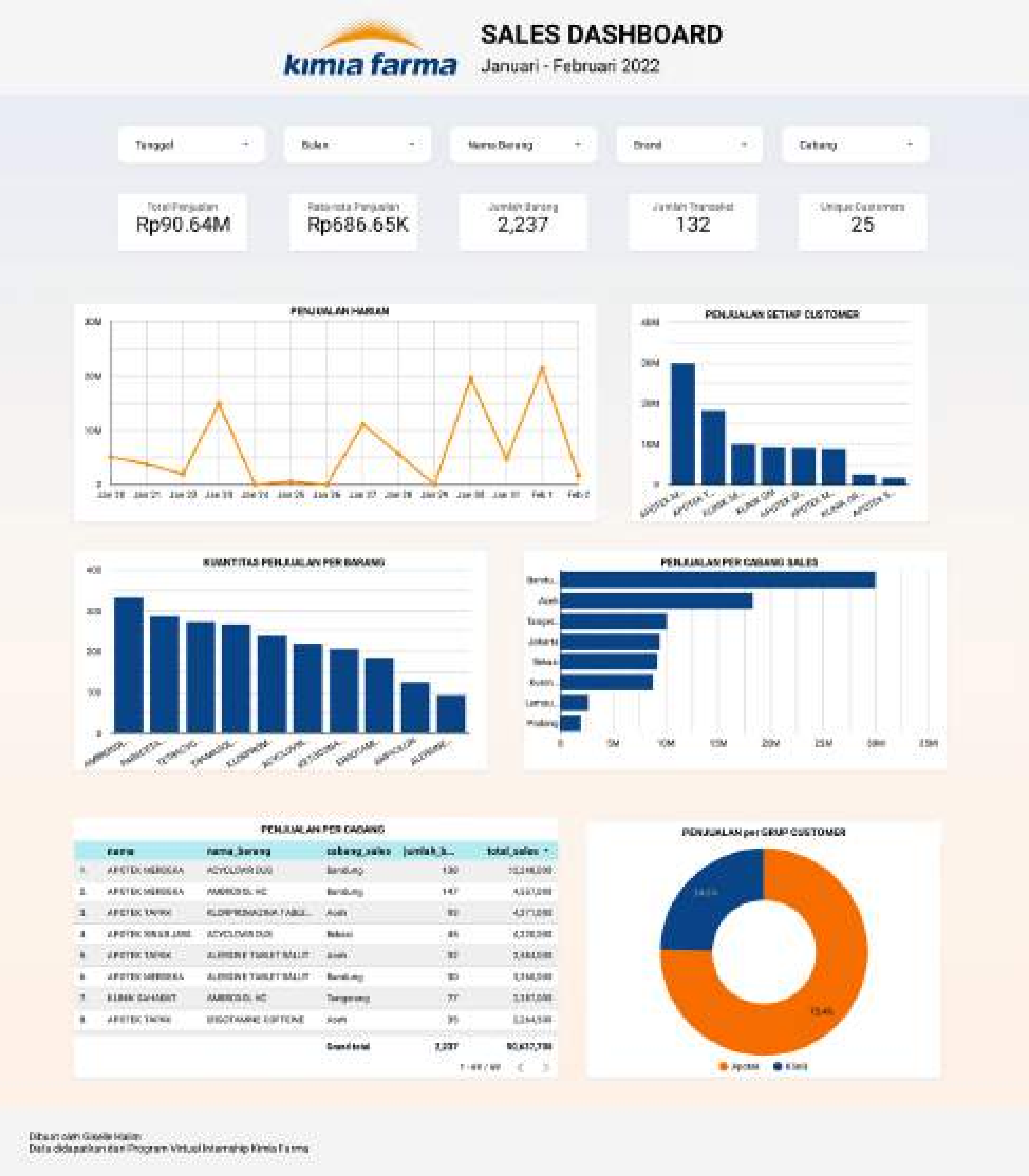
Source: IDF Diabetes Atlas

Diabetes imposes a significant financial burden on healthcare systems. **Rising diabetes costs can deplete health budgets, affecting other essential services.** This signifies the urgent need for effective diabetes management to prevent economic strain and ensure equitable healthcare access.

A comprehensive analysis of diabetes cases in ASEAN was conducted using SAP Analytics Cloud. This project focused on understanding the rising prevalence of diabetes, its associated complications, mortality rates, risk factors, and the significant economic burden it imposes on the region. By aligning with Sustainable Development Goal 3 (Good Health and Wellbeing), the analysis sought to identify effective solutions. In addition to sector-specific recommendations, a gamified app was conceptualized to promote reduced sugar consumption and contribute to mitigating the diabetes epidemic. This project was selected to represent our team in the national finals of the ASEAN Data Science Explorers 2024 Competition, underscoring its significance and potential impact.

FULL SLIDE

MEDICINE SALES DASHBOARD



kimia farma

SALES DASHBOARD

Januari - Februari 2022

Tanggal | Bulan | Nama Barang | Brand | Cabang

Total Penjualan Rp90.64M

Rata-rata Penjualan Rp686.65K

Jumlah Barang 2,237

Jumlah Transaksi 132

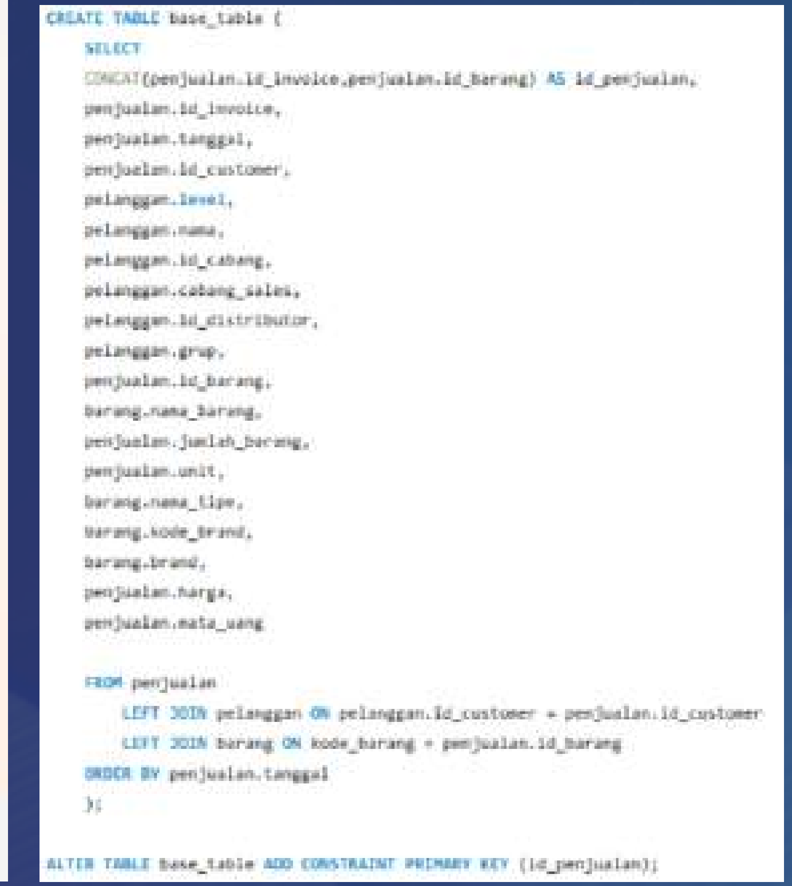
Unique Customers 25

PENJUALAN HARIAN
PENJUALAN SETIAP CUSTOMER
KUANTITAS PENJUALAN PER BARANG
PENJUALAN PER CABANG SALES
PENJUALAN PER CABANG
PENJUALAN per GRUP CUSTOMER
Apotek
Klinik

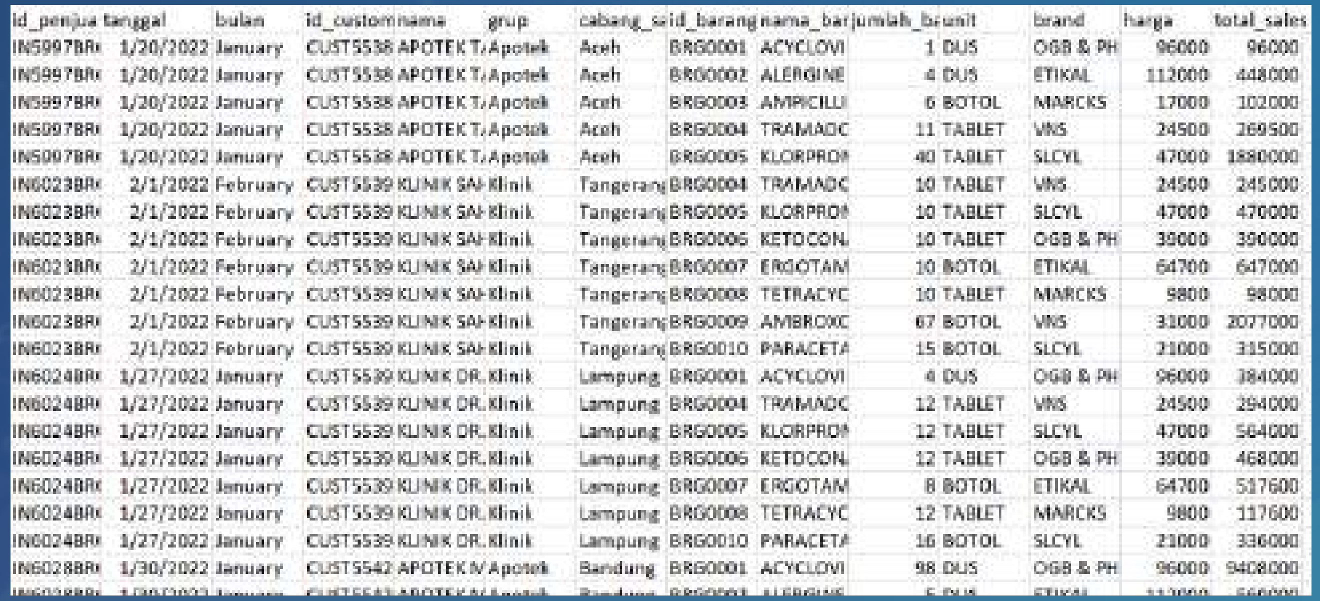
Dibuat oleh Giselle Halim

Data didapatkan dari Program Virtual Internship Kimia Farma

Leveraged SQL to cleanse and structure data for a comprehensive dashboard analysis of medicine sales at Kimia Farma. Developed a dashboard with Looker to visualize sales trends, total revenue, and product-level performance over a two-week period. The dashboard facilitates quick analysis and data-driven decision-making for optimizing inventory and sales strategies.



```
CREATE TABLE base_table (
    SELECT
    CONCAT(penjualan.id_invoice,penjualan.id_barang) AS id_penjualan,
    penjualan.id_invoice,
    penjualan.tanggal,
    penjualan.id_customer,
    pelanggan.level,
    pelanggan.nama,
    pelanggan.id_cabang,
    pelanggan.cabang_sales,
    pelanggan.id_distributor,
    pelanggan.grup,
    penjualan.id_barang,
    barang.nama_barang,
    penjualan.jumlah_barang,
    penjualan.unit,
    barang.nama_tipe,
    barang.kode_brand,
    barang.brand,
    penjualan.harga,
    penjualan.mata_uang

    FROM penjualan
        LEFT JOIN pelanggan ON pelanggan.id_customer = penjualan.id_customer
        LEFT JOIN barang ON kode_barang = penjualan.id_barang
    ORDER BY penjualan.tanggal
    );

ALTER TABLE base_table ADD CONSTRAINT PRIMARY KEY (id_penjualan);
```

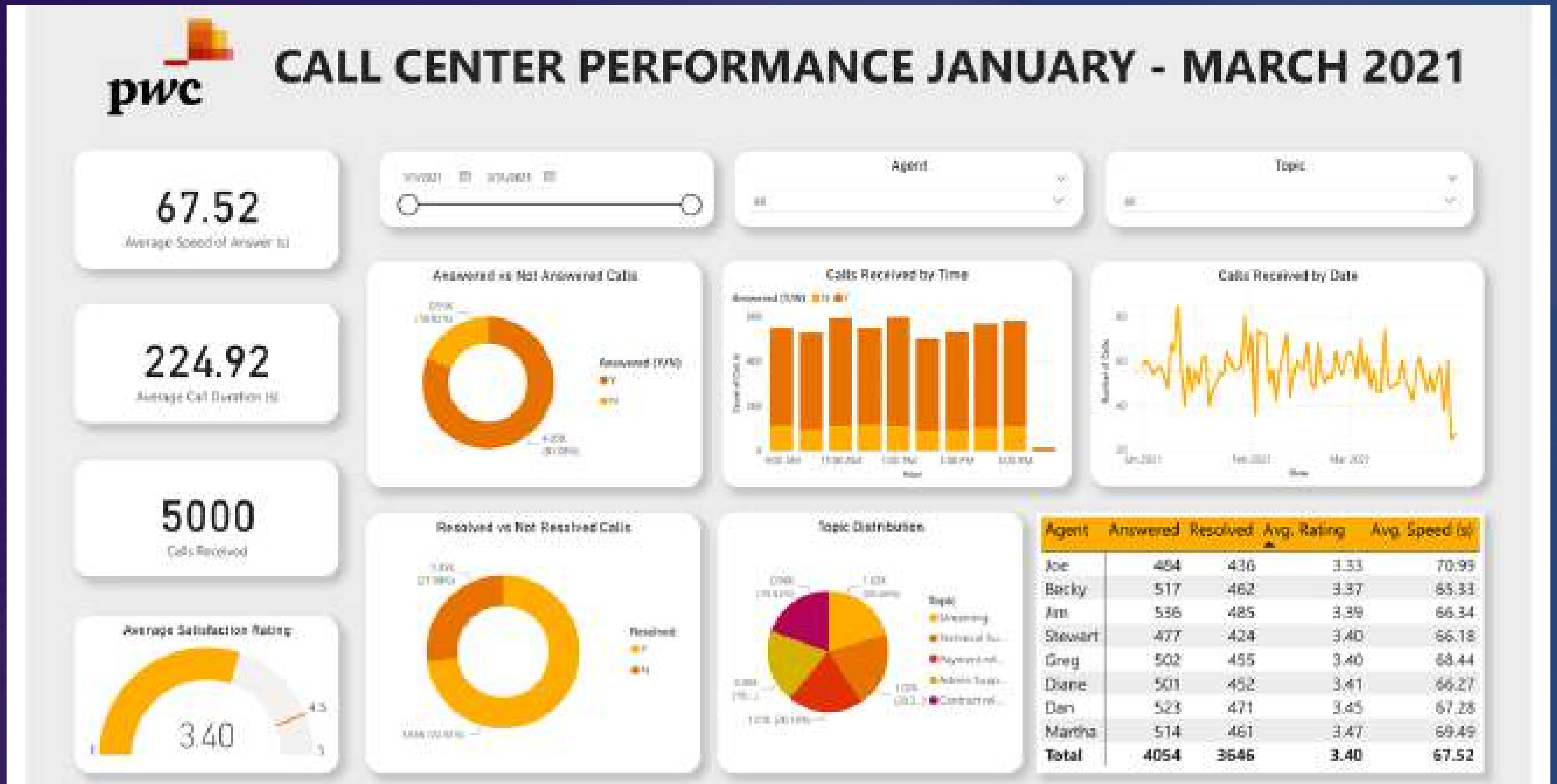


id_penjua	tanggal	bulan	id_custom	nama	grup	cabang_sa	id_barang	nama_bar	jumlah_b	unit	brand	harga	total_sales
IN5997BR	1/20/2022	January	CUST5538	APOTEK T.	Apotek	Aceh	BRG0001	ACYCLOVI	1	DUS	OGB & PH	96000	96000
IN5997BR	1/20/2022	January	CUST5538	APOTEK T.	Apotek	Aceh	BRG0002	ALERGINE	4	DUS	ETIKAL	112000	448000
IN5997BR	1/20/2022	January	CUST5538	APOTEK T.	Apotek	Aceh	BRG0003	AMPICILLI	6	BOTOL	MARCKS	17000	102000
IN5997BR	1/20/2022	January	CUST5538	APOTEK T.	Apotek	Aceh	BRG0004	TRAMADC	11	TABLET	VNS	24500	269500
IN5997BR	1/20/2022	January	CUST5538	APOTEK T.	Apotek	Aceh	BRG0005	KLORPROM	40	TABLET	SLCYL	47000	1880000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SAH	Klinik	Tangerang	BRG0004	TRAMADC	10	TABLET	VNS	24500	245000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SAH	Klinik	Tangerang	BRG0005	KLORPROM	10	TABLET	SLCYL	47000	470000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SAH	Klinik	Tangerang	BRG0006	KETOCON.	10	TABLET	OGB & PH	39000	390000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SAH	Klinik	Tangerang	BRG0007	ERGOTAM	10	BOTOL	ETIKAL	64700	647000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SAH	Klinik	Tangerang	BRG0008	TETRACYC	10	TABLET	MARCKS	9800	98000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SAH	Klinik	Tangerang	BRG0009	AMBROXC	67	BOTOL	VNS	31000	2077000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SAH	Klinik	Tangerang	BRG0010	PARACETA	15	BOTOL	SLCYL	21000	315000
IN6024BR	1/27/2022	January	CUST5539	KLINIK DR.	Klinik	Lampung	BRG0001	ACYCLOVI	4	DUS	OGB & PH	96000	384000
IN6024BR	1/27/2022	January	CUST5539	KLINIK DR.	Klinik	Lampung	BRG0004	TRAMADC	12	TABLET	VNS	24500	294000
IN6024BR	1/27/2022	January	CUST5539	KLINIK DR.	Klinik	Lampung	BRG0005	KLORPROM	12	TABLET	SLCYL	47000	564000
IN6024BR	1/27/2022	January	CUST5539	KLINIK DR.	Klinik	Lampung	BRG0006	KETOCON.	12	TABLET	OGB & PH	39000	468000
IN6024BR	1/27/2022	January	CUST5539	KLINIK DR.	Klinik	Lampung	BRG0007	ERGOTAM	8	BOTOL	ETIKAL	64700	517600
IN6024BR	1/27/2022	January	CUST5539	KLINIK DR.	Klinik	Lampung	BRG0008	TETRACYC	12	TABLET	MARCKS	9800	117600
IN6024BR	1/27/2022	January	CUST5539	KLINIK DR.	Klinik	Lampung	BRG0010	PARACETA	16	BOTOL	SLCYL	21000	336000
IN6028BR	1/30/2022	January	CUST5542	APOTEK IV	Apotek	Bandung	BRG0001	ACYCLOVI	98	DUS	OGB & PH	96000	9408000

FULL PROJECT HERE

LOOKER DASHBOARD

CALL CENTER DASHBOARD



CALL CENTER PERFORMANCE JANUARY - MARCH 2021

67.52 Average Speed of Answer (s)

224.92 Average Call Duration (s)

5000 Calls Received

pwc
Agent
Topic
Answered vs Not Answered Calls
Calls Received by Time
Calls Received by Date
Average Satisfaction Rating
3.40
Resolved vs Not Resolved Calls
Topic Distribution

Agent	Answered	Resolved	Avg. Rating	Avg. Speed (s)
Joe	484	436	3.33	70.99
Becky	517	462	3.37	65.33
Jim	536	485	3.39	66.34
Stewart	477	424	3.40	66.18
Greg	502	455	3.40	68.44
Diane	501	452	3.41	66.27
Dan	523	471	3.45	67.28
Martha	514	461	3.47	69.49
Total	**4054**	**3646**	**3.40**	**67.52**

In today's saturated telecom market, where providers bombard customers with claims of "better price" and "best service," a clear understanding of customer needs is crucial. This Power BI dashboard empowers a major telecom company to cut through the noise. By tracking KPIs like overall customer satisfaction, call answer rates, and call duration, the dashboard provides actionable insights to improve customer experience and optimize call center operations.

INSIGHTS

CALL CENTER PERFORMANCE

- Average Speed of Answer (ASA) is longer than industry standard at **67.52 seconds**.
- Average Handle Time (AHT) is **224.92 seconds or 3.7 minutes**.
- Call abandonment rate is **19%**, suggesting opportunities to improve service levels.
- Call resolution rate is **73%**, indicating room for improvement in first-call resolution.

CALL VOLUME PATTERNS

- Call volume peaks at **11 AM, 1 PM, and 5 PM**, requiring optimized staffing during these times.
- An unusually high call volume was recorded on **January 11, 2021**.

PROBLEMS

01 Long Average Speed of Answer (ASA)

The current ASA of 67.52 seconds significantly exceeds the industry standard of 20-30 seconds, indicating a need to improve call handling processes.

02 Customer Satisfaction Gap

The current average customer satisfaction rating of 3.4 falls short of the desired 4.5, highlighting a need to enhance customer experience through improved service quality and resolution.

03 Call Volume Fluctuations

Call volume peaks at 11 AM, 1 PM, and 5 PM, resulting in potential service disruptions. Optimized staffing is required to manage these peak periods effectively.

SOLUTIONS

01 Speed up Answer Times

Focus on reducing average speed of answer by analyzing peak call times and optimizing staffing levels during these periods. Identify and address root causes of long wait times, such as system slowdowns or agent training gaps.

02 Enhance Customer Satisfaction

Implement targeted training for agents, especially for Joe, who has the lowest rating. Conduct regular customer satisfaction surveys to pinpoint areas for improvement. Consider offering incentives for high-performing agents.

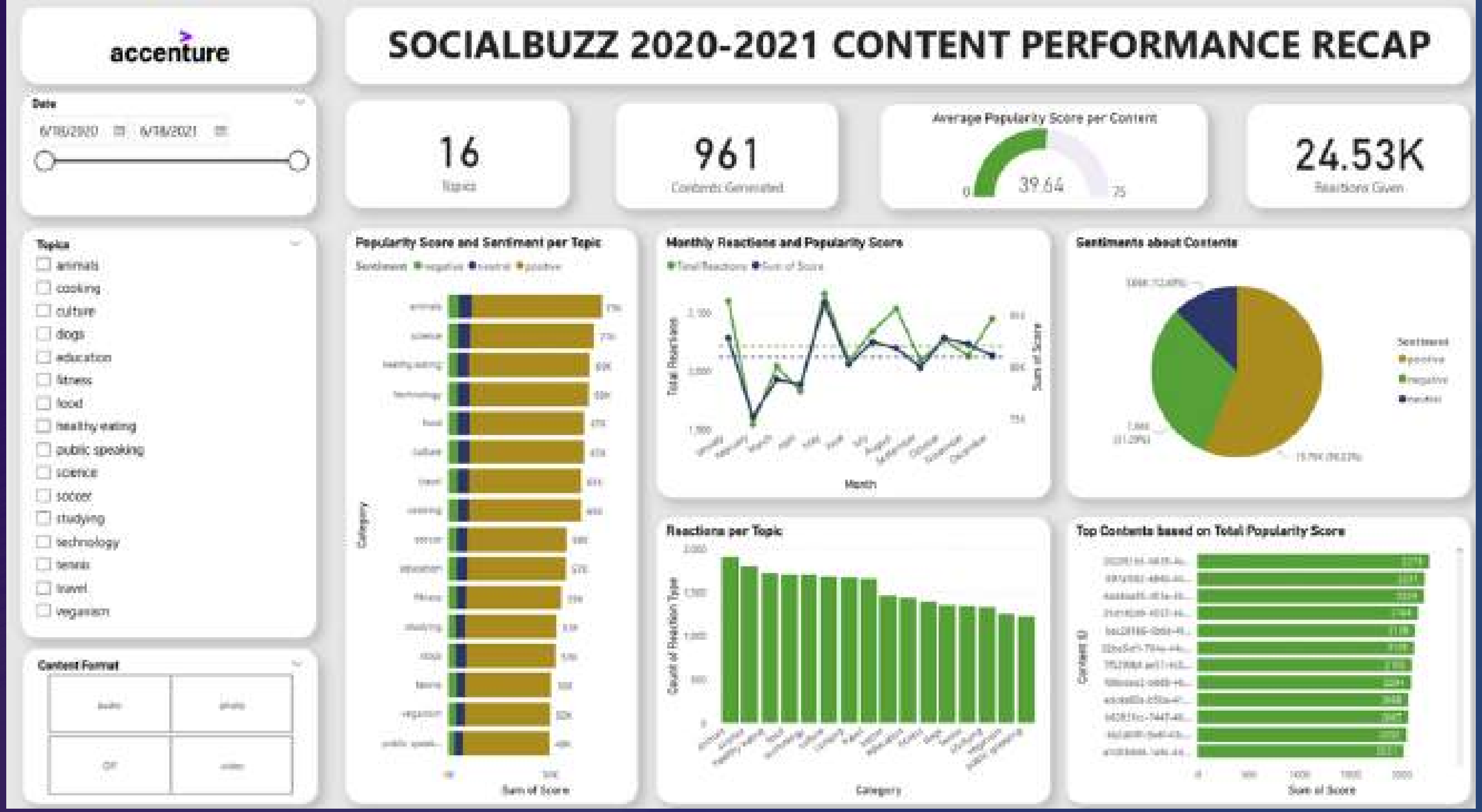
03 Optimize Call Handling

Analyze call duration data to identify reasons for particularly short or long calls. Provide agents with tools and knowledge to resolve issues efficiently. Consider implementing call routing based on issue type to direct calls to specialized agents.

FULL PROJECT HERE

FULL SLIDE

SOCIAL MEDIA CONTENT ANALYSIS



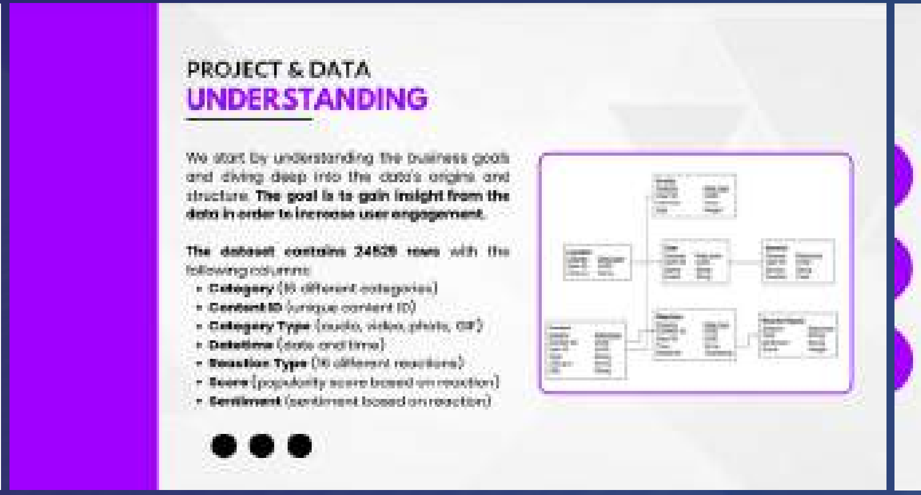
This data analysis project, conducted for SocialBuzz, a leading social media and content creation firm, aimed to optimize their content strategy using Power BI. By analyzing vast amounts of social media data, the project delivered actionable insights into content performance, audience engagement, and trends. The analysis focused on identifying the top 5 content categories driving the most engagement, providing recommendations for content optimization, and uncovering opportunities for further growth.

PROBLEM

Over 100,000 contents per day ranging from text, images, videos, and GIFs.

More than 36.5 million contents per year. This signifies rapid growth and massive data volume.

How to use the data to increase user engagement? Analysis of the top 5 most popular content categories, top contents, and monthly content performance.



PROJECT & DATA UNDERSTANDING

We start by understanding the business goals and diving deep into the data's origins and structure. **The goal is to gain insight from the data in order to increase user engagement.**

The dataset contains 24528 rows with the following columns:

- **Category** (16 different categories)
- **Content ID** (unique content ID)
- **Category Type** (audio, video, photo, GIF)
- **Datetime** (date and time)
- **Reaction Type** (16 different reactions)
- **Score** (popularity score based on reaction)
- **Sentiment** (sentiment based on reaction)

SUMMARY

The most popular category, 'animals' has 1897 reactions and 74965 total popularity score. The other top categories are "science", "healthy eating", "technology", and "food".

Animals and science are the two most popular categories of content, showing that **people enjoy 'real-life' and 'factual' content the most.**

May 2021 was the month with the most post reactions with a total of [illegible] reactions, followed by January 2021 and August 2020.

Food is a common theme with the top 5 categories with 'Healthy Eating' ranking the highest. You could use this insight to create a campaign and work with healthy eating brands to boost user engagement.

Most of the content gains positive sentiment with an average popularity score of 39.64. This means that the post is generally positive.

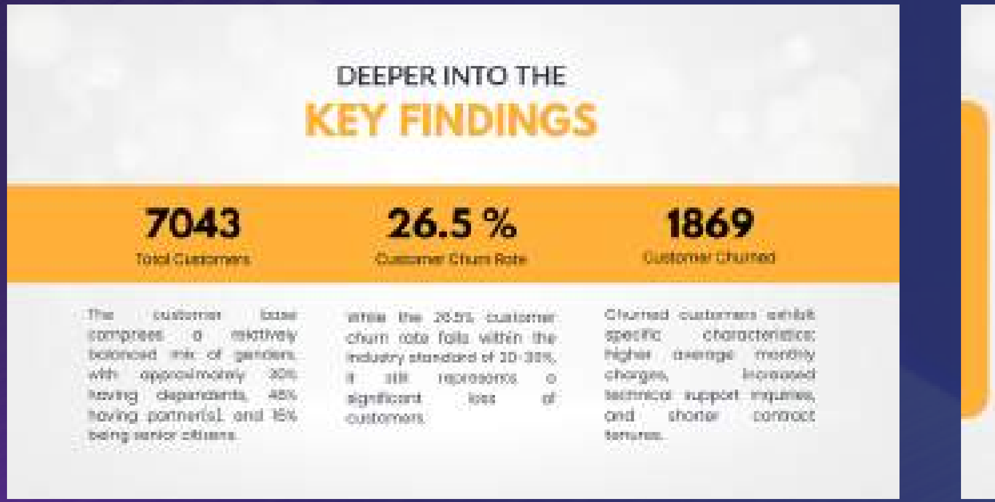
This ad-hoc analysis is insightful but it's time to **take this analysis into large scale production for real-time understanding of the business.**

FULL PROJECT HERE

FULL SLIDE

CUSTOMER CHURN DASHBOARD

This data analysis project focused on understanding customer churn within a telecom company. By leveraging Power BI, a comprehensive dashboard was developed to visualize customer demographics, service usage patterns, and other relevant factors contributing to churn. While the overall churn rate of 26.5% fell within industry standards, it highlighted the need for proactive retention strategies. The dashboard provided valuable insights to identify at-risk customers and implement targeted interventions to mitigate churn and enhance customer satisfaction.



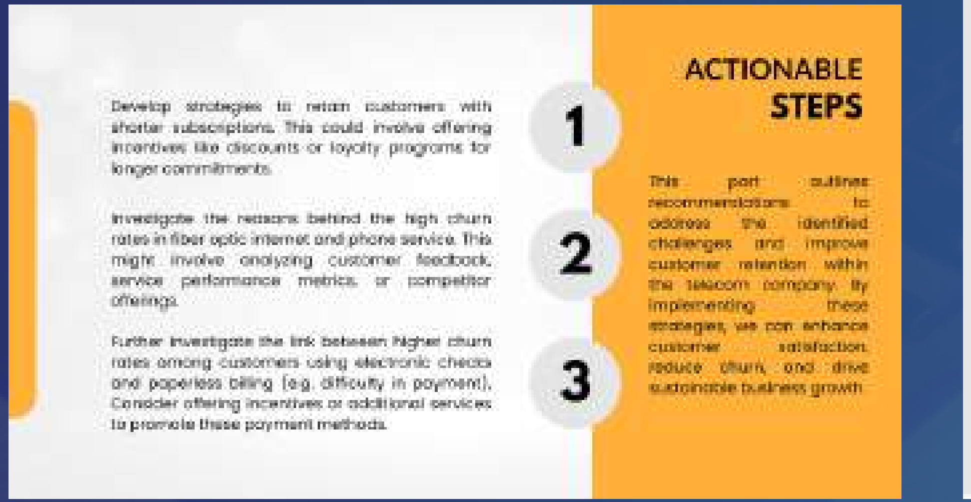
DEEPER INTO THE
KEY FINDINGS

7043	26.5 %	1869
Total Customers	Customer Churn Rate	Customer Churned

The customer base comprises a relatively balanced mix of genders, with approximately 30% having dependents, 48% having partner(s), and 16% being senior citizens.

While the 26.5% customer churn rate falls within the industry standard of 20-30%, it still represents a significant loss of customers.

Churned customers exhibit specific characteristics: higher average monthly charges, increased technical support inquiries, and shorter contract tenures.



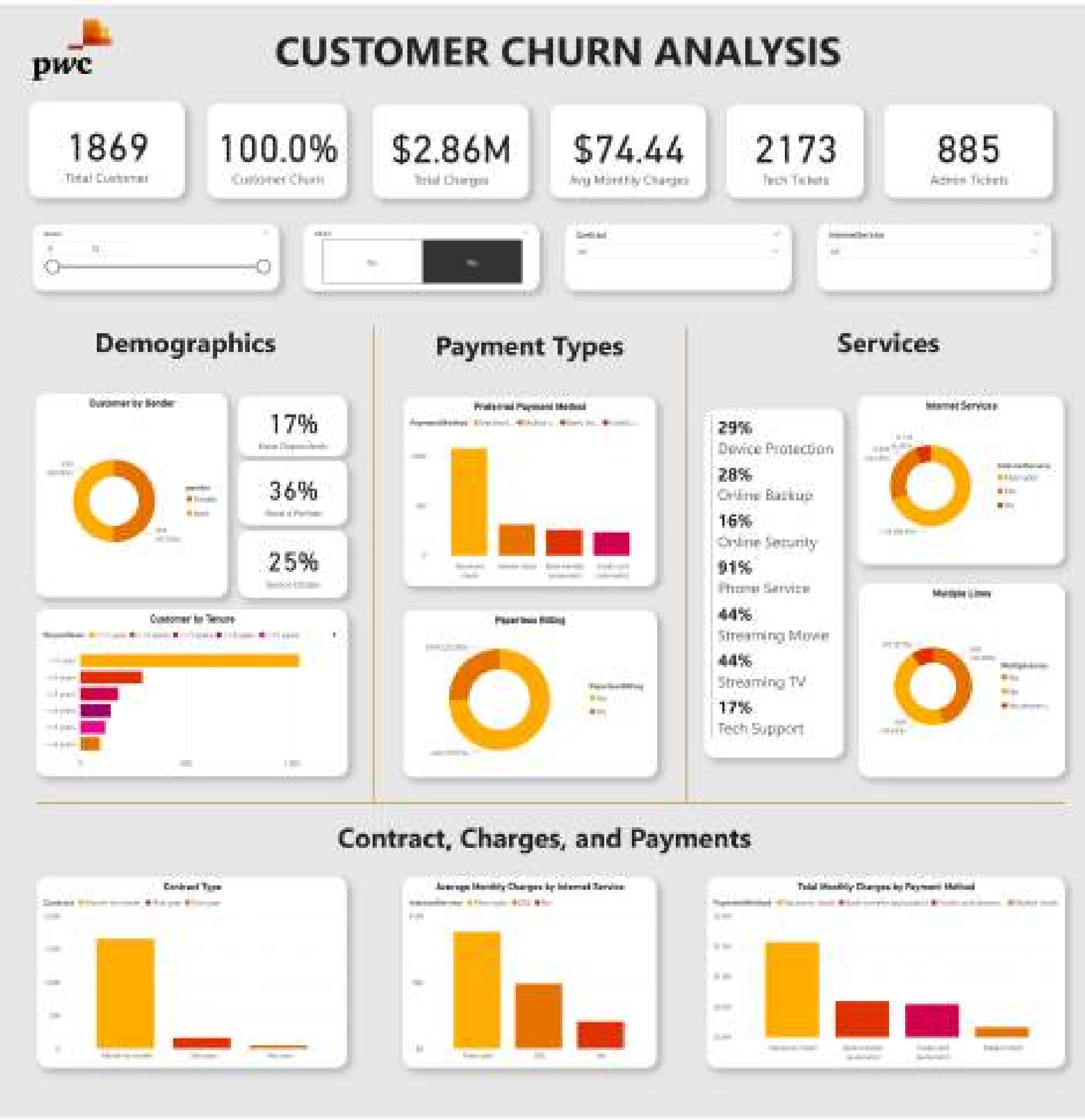
ACTIONABLE STEPS

This part outlines recommendations to address the identified challenges and improve customer retention within the telecom company. By implementing these strategies, we can enhance customer satisfaction, reduce churn, and drive sustainable business growth.

1. Develop strategies to retain customers with shorter subscriptions. This could involve offering incentives like discounts or loyalty programs for longer commitments.

2. Investigate the reasons behind the high churn rates in fiber optic internet and phone service. This might involve analyzing customer feedback, service performance metrics, or competitor offerings.

3. Further investigate the link between higher churn rates among customers using electronic checks and paperless billing (e.g. difficulty in payment). Consider offering incentives or additional services to promote these payment methods.



CUSTOMER CHURN ANALYSIS

1869	100.0%	$2.86M	$74.44	2173	885
Total Customer	Customer Churn	Total Charges	Avg Monthly Charges	Tech Tickets	Admin Tickets

Demographics

17%	36%	25%
[illegible]	[illegible]	[illegible]

Payment Types

Services

29%	28%	16%	91%	44%	44%	17%
Device Protection	Online Backup	Online Security	Phone Service	Streaming Movie	Streaming TV	Tech Support

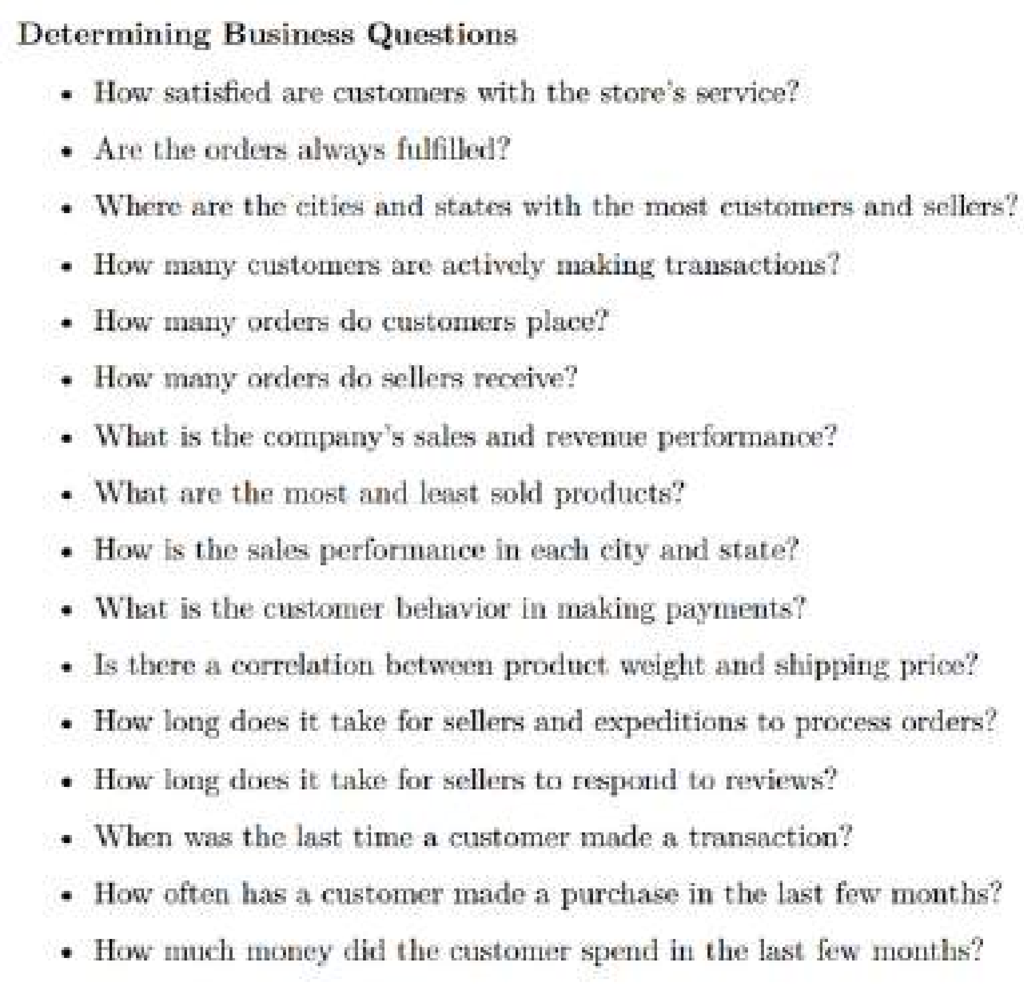
Contract, Charges, and Payments

FULL PROJECT HERE

FULL SLIDE

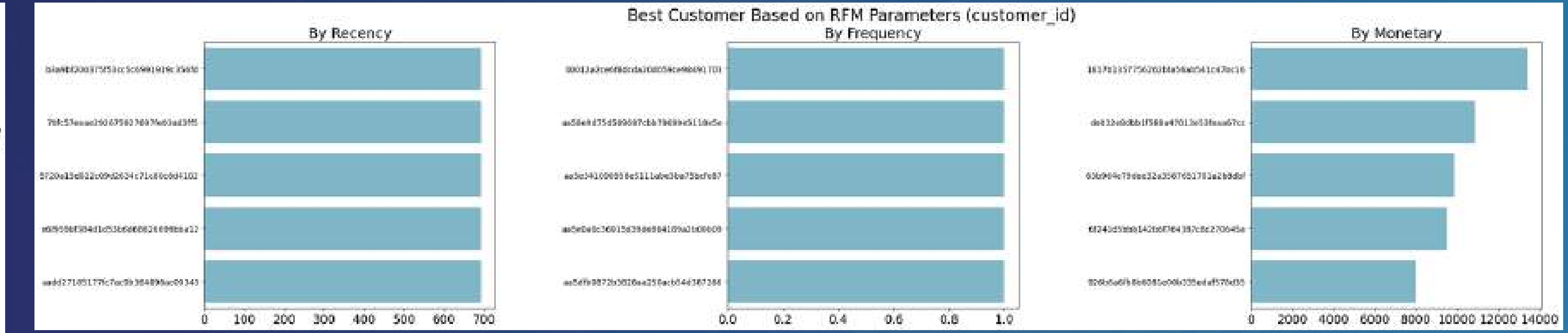
E-COMMERCE SALES ANALYSIS

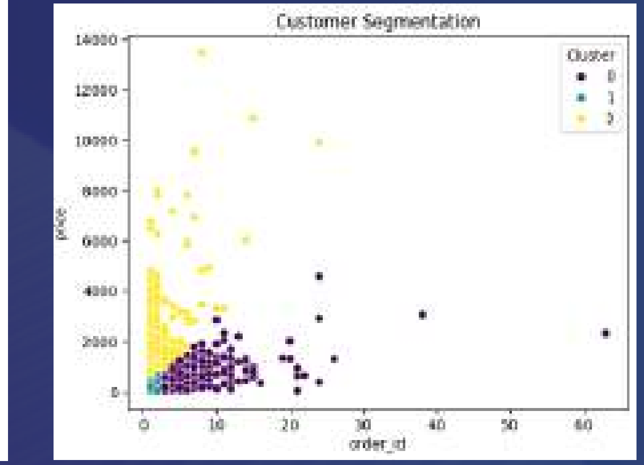
Conducted data analysis for an e-commerce platform using Python to address key business questions. The analysis included calculating total sales and profit, evaluating monthly sales trends, identifying preferred payment methods, and performing RFM (Recency, Frequency, Monetary) analysis. K-Means Clustering was implemented to segment customers, enabling targeted marketing strategies. This provided insights into sales performance, customer behavior, and preferences, aiding in strategic decision-making. A web dashboard built with Streamlit provided a simple summary of the transaction data.

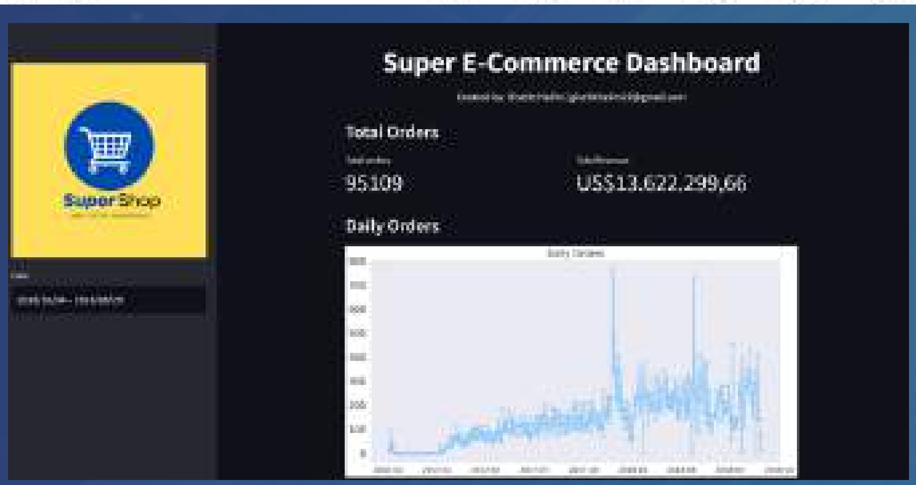


Determining Business Questions

- How satisfied are customers with the store's service?
- Are the orders always fulfilled?
- Where are the cities and states with the most customers and sellers?
- How many customers are actively making transactions?
- How many orders do customers place?
- How many orders do sellers receive?
- What is the company's sales and revenue performance?
- What are the most and least sold products?
- How is the sales performance in each city and state?
- What is the customer behavior in making payments?
- Is there a correlation between product weight and shipping price?
- How long does it take for sellers and expeditions to process orders?
- How long does it take for sellers to respond to reviews?
- When was the last time a customer made a transaction?
- How often has a customer made a purchase in the last few months?
- How much money did the customer spend in the last few months?






Super E-Commerce Dashboard

Total Orders

95109

US$13.622.299,66

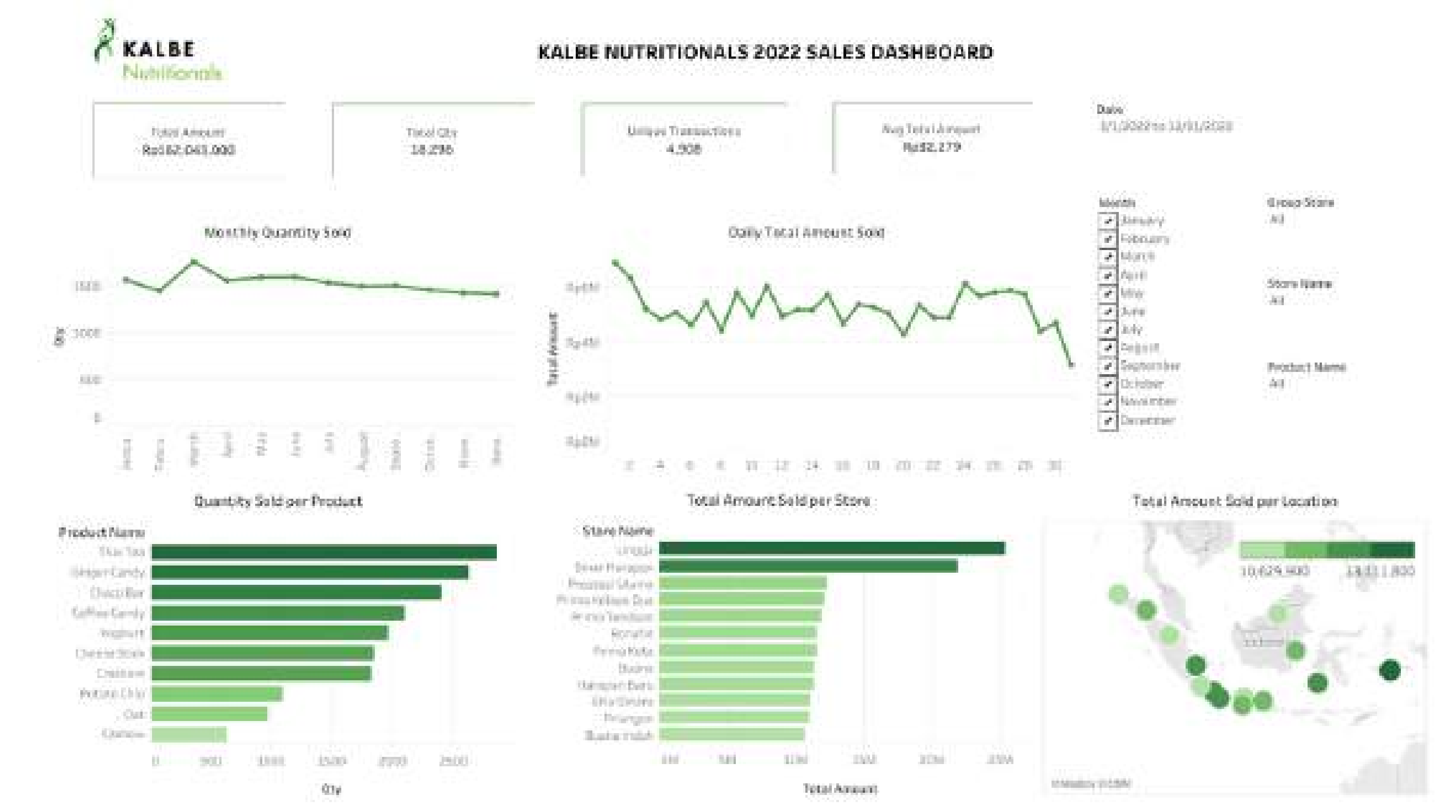
Daily Orders

FULL PROJECT HERE

STREAMLIT APP

HEALTH PRODUCT SALES ANALYSIS

Leveraged SQL to analyze customer and store data, gaining valuable insights for business improvement. Python was employed for exploratory data analysis, while a Tableau dashboard provided a comprehensive overview of health product sales performance, including metrics like total sales, revenue, and sales trends. The dashboard featured visualizations of sales patterns, product popularity, and regional distribution, enabling swift identification of top-performing areas and products. K-Means Clustering was implemented to segment customers, enabling targeted marketing strategies.



KALBE NUTRITIONALS 2022 SALES DASHBOARD

Total Amount	Total Qty	Unique Transactions	Avg Total Amount
Rp162.043.000	18.296	4.908	Rp32.279

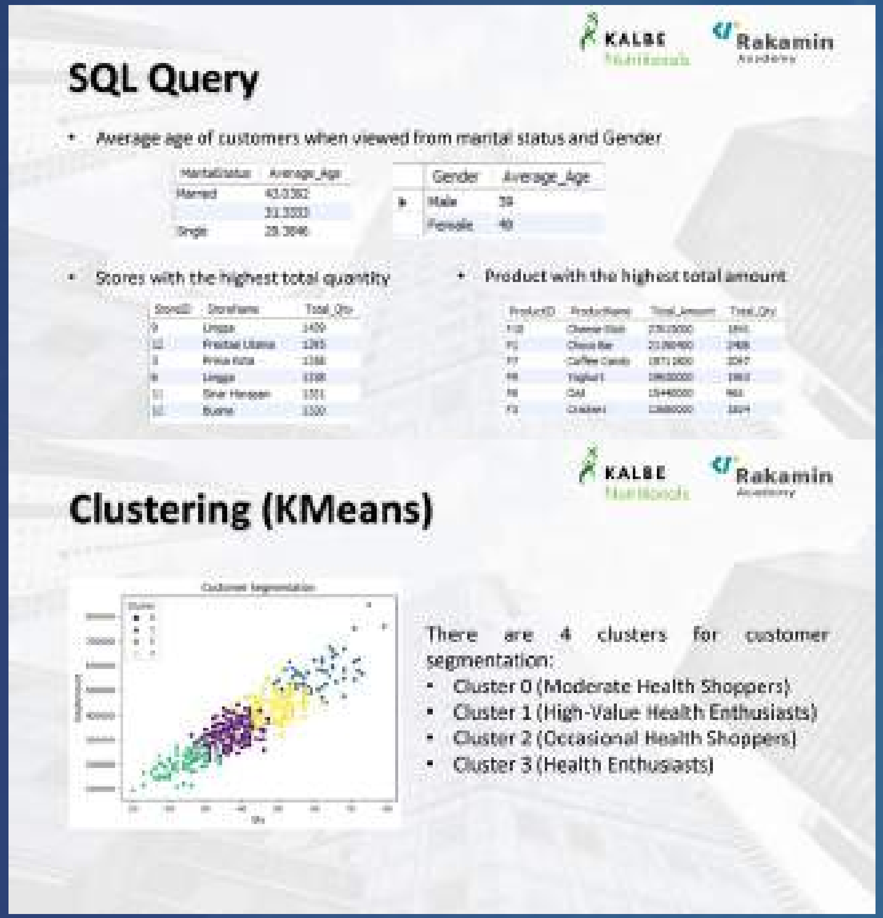
Date: 1/1/2022 to 12/31/2022

Month: ☑ January ☑ February ☑ March ☑ April ☑ May ☑ June ☑ July ☑ August ☑ September ☑ October ☑ November ☑ December

Group Store: All

Store Name: All

Product Name: All



SQL Query

- Average age of customers when viewed from marital status and Gender

MaritalStatus	Average_Age
Married	43.0382
	31.3333
Single	29.3846

Gender	Average_Age
Male	39
Female	40

- Stores with the highest total quantity

StoreID	StoreName	Total_Qty
9	Lingga	1409
12	Prestasi Utama	1295
3	Prima Kota	1188
6	Lingga	1108
11	Sinar Harapan	1101
10	Buana	1100

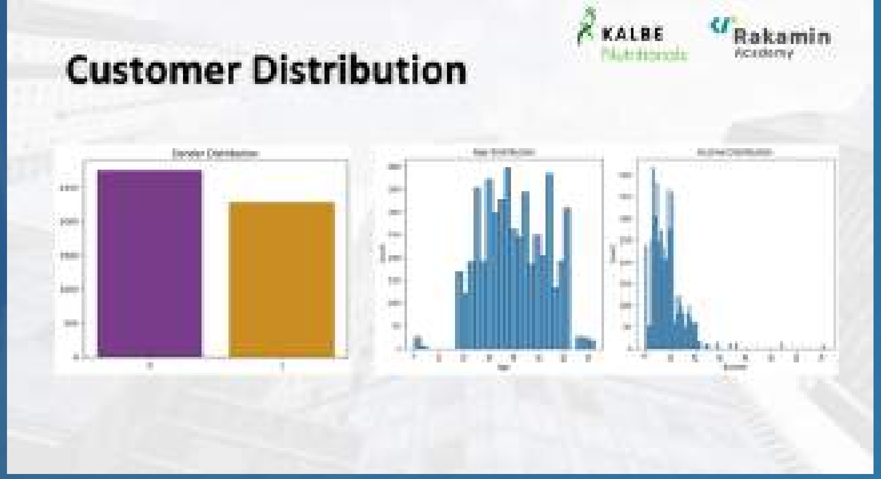
- Product with the highest total amount

ProductID	ProductName	Total_Amount	Total_Qty
F10	Cheese Stick	27615000	1841
F5	Choco Bar	21198400	2408
F7	Coffee Candy	18711000	2047
F8	Yoghurt	19620000	1962
F6	Oat	15440000	962
F3	Crackers	13680000	1824

Clustering (KMeans)

There are 4 clusters for customer segmentation:

- Cluster 0 (Moderate Health Shoppers)
- Cluster 1 (High-Value Health Enthusiasts)
- Cluster 2 (Occasional Health Shoppers)
- Cluster 3 (Health Enthusiasts)



Customer Distribution

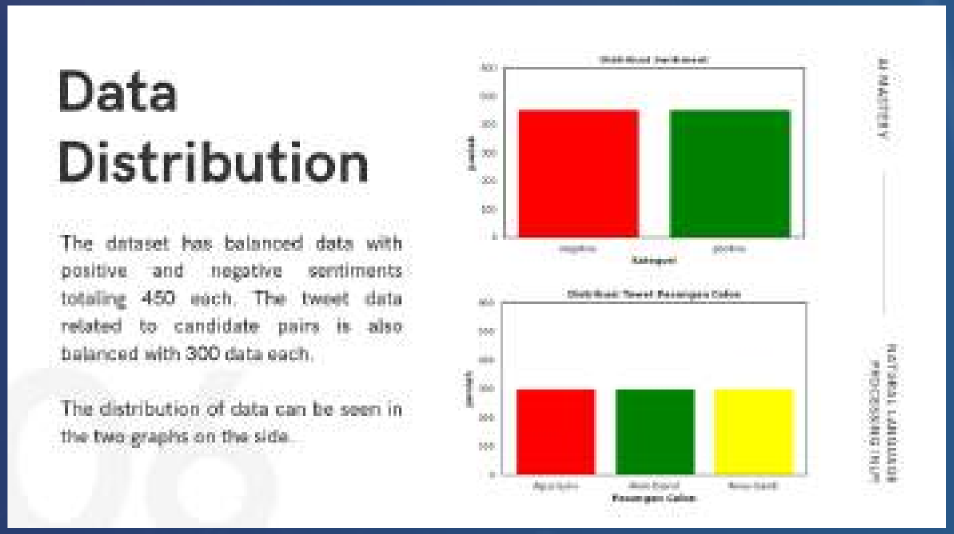
FULL PROJECT HERE

TABLEAU DASHBOARD

FULL SLIDE

2017 JAKARTA LOCAL ELECTION SENTIMENT ANALYSIS

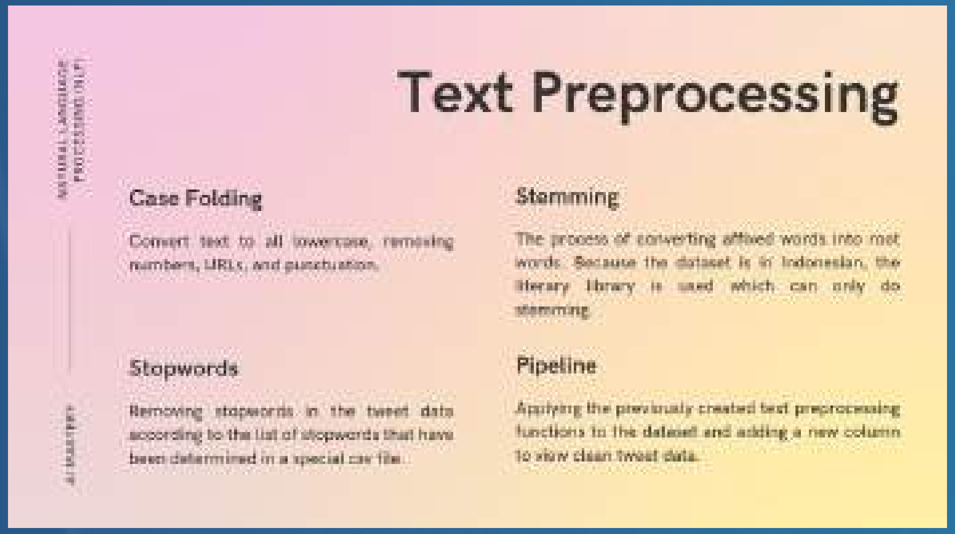
This project employed a Bernoulli Naive Bayes classifier to conduct sentiment analysis on tweets related to the 2017 DKI Jakarta Local Leader Election. Data preprocessing like case folding, stemming, and removing stopwords were done. Feature extraction techniques were implemented to optimize model performance. The model effectively analyzed public sentiment towards the three candidates, providing valuable insights into voter preferences during the election campaign. With an accuracy rate of 88%, the model demonstrated its ability to accurately gauge public opinion based on social media data.



Data Distribution

The dataset has balanced data with positive and negative sentiments totaling 450 each. The tweet data related to candidate pairs is also balanced with 300 data each.

The distribution of data can be seen in the two graphs on the side.



Text Preprocessing

Case Folding

Convert text to all lowercase, removing numbers, URLs, and punctuation.

Stemming

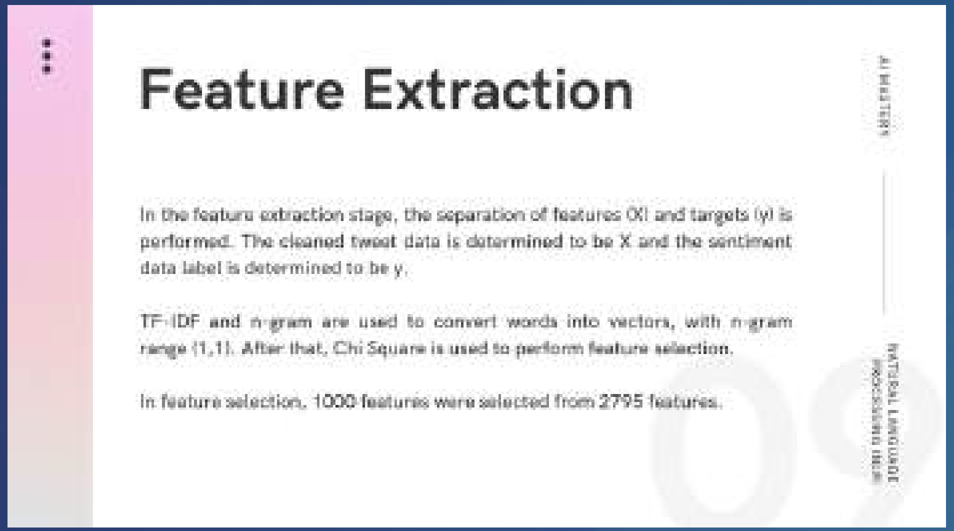
The process of converting affixed words into root words. Because the dataset is in Indonesian, the literary library is used which can only do stemming.

Stopwords

Removing stopwords in the tweet data according to the list of stopwords that have been determined in a special csv file.

Pipeline

Applying the previously created text preprocessing functions to the dataset and adding a new column to view clean tweet data.

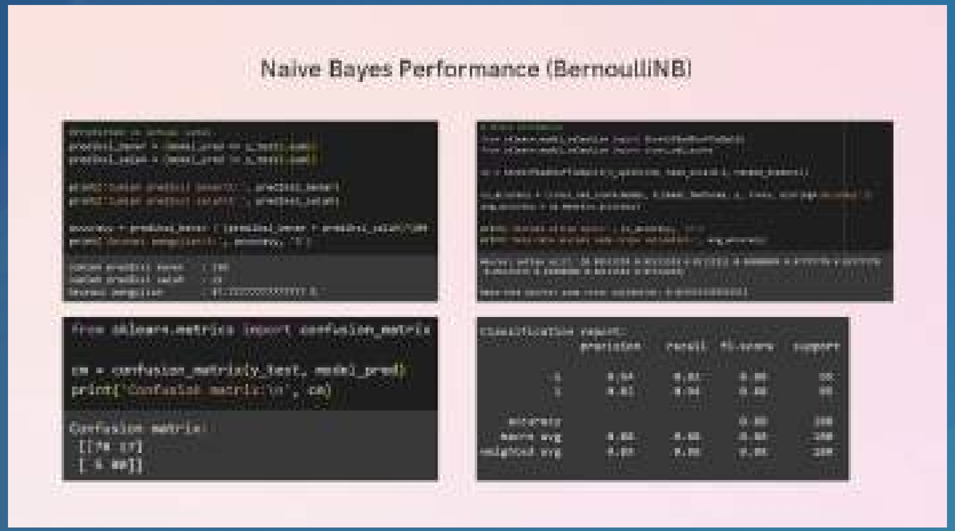


Feature Extraction

In the feature extraction stage, the separation of features (X) and targets (y) is performed. The cleaned tweet data is determined to be X and the sentiment data label is determined to be y.

TF-IDF and n-gram are used to convert words into vectors, with n-gram range (1,1). After that, Chi Square is used to perform feature selection.

In feature selection, 1000 features were selected from 2795 features.



Naive Bayes Performance (BernoulliNB)

```
from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, model_pred)
print('Confusion matrix:\n', cm)
```

```
Confusion matrix:
[illegible]
```

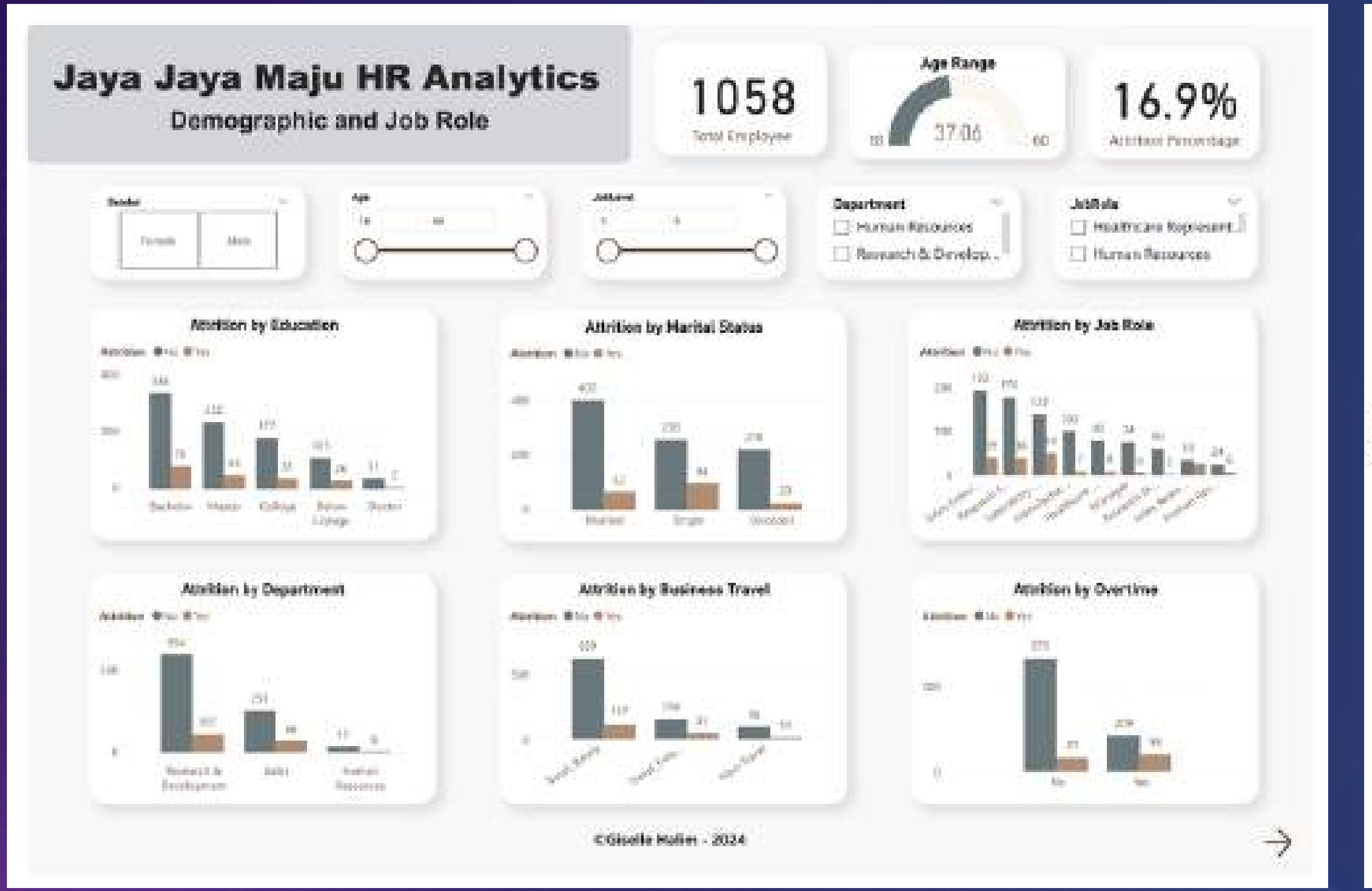
```
Classification report:
              precision    recall  f1-score   support
[illegible]
    accuracy
   macro avg
weighted avg
```

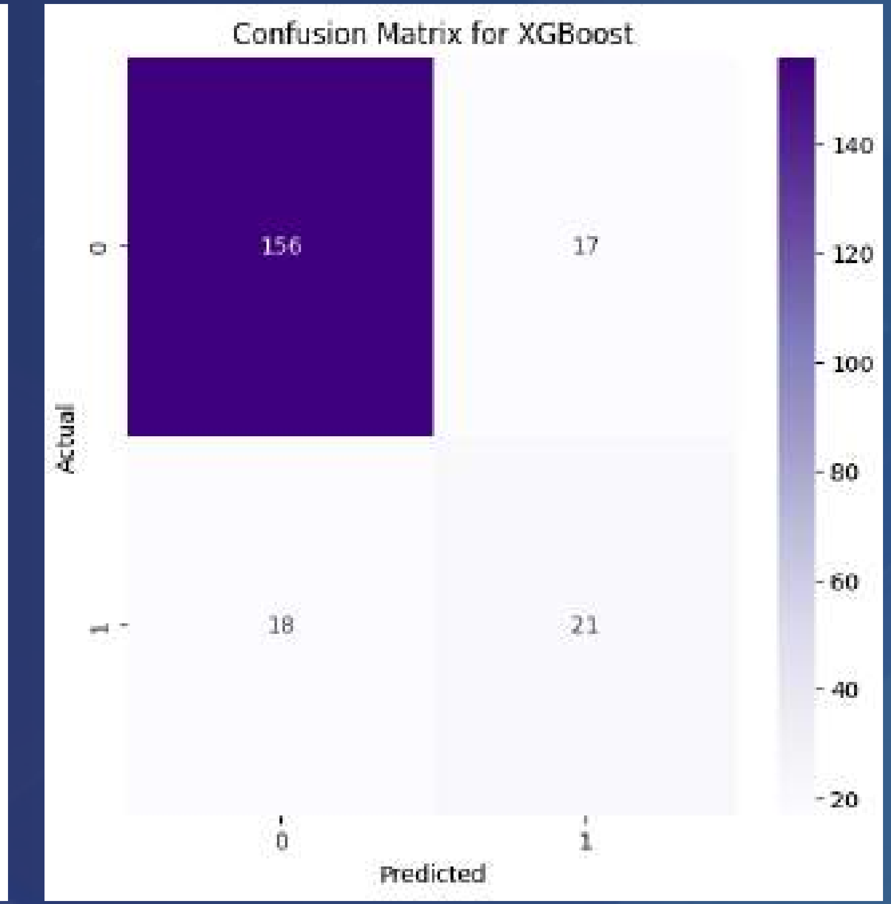
FULL PROJECT HERE

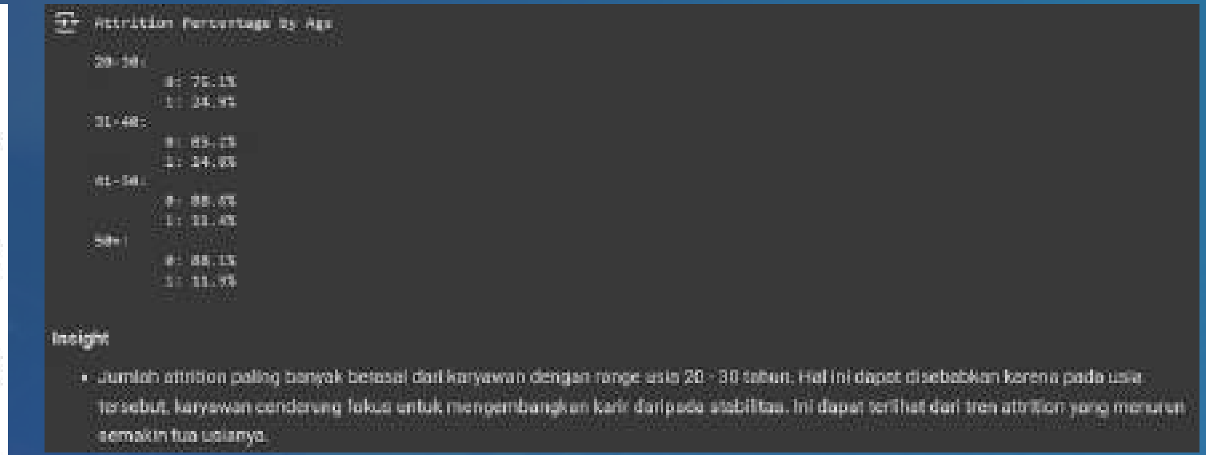
FULL SLIDE

EMPLOYEE ATTRITION ANALYSIS

Leveraged Python and machine learning to analyze over 1,000 employee data and predict attrition rates by examining factors like job role, department, business travel, and satisfaction. A predictive model built with XGBoost achieved 83% accuracy in predicting employee departures. PowerBI dashboards visualized the findings to inform data-driven retention strategies.



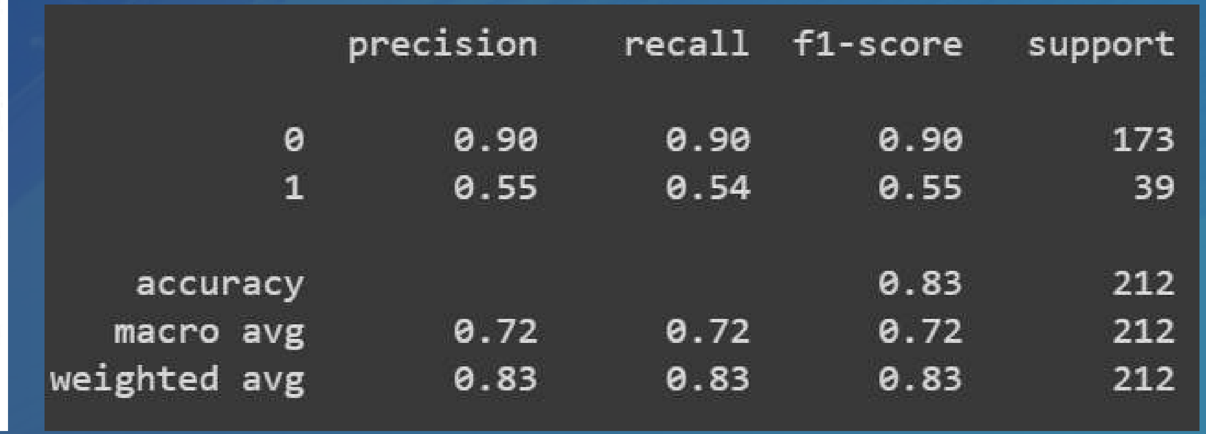




```
Attrition Percentage by Age
  20-30:
        0: 75.1%
        1: 24.9%
  31-40:
        0: 85.2%
        1: 14.8%
  41-50:
        0: 88.6%
        1: 11.4%
  50+:
        0: 88.1%
        1: 11.9%
```

Insight

- Jumlah attrition paling banyak berasal dari karyawan dengan range usia 20 - 30 tahun. Hal ini dapat disebabkan karena pada usia tersebut, karyawan cenderung fokus untuk mengembangkan karir daripada stabilitas. Ini dapat terlihat dari tren attrition yang menurun semakin tua usianya.

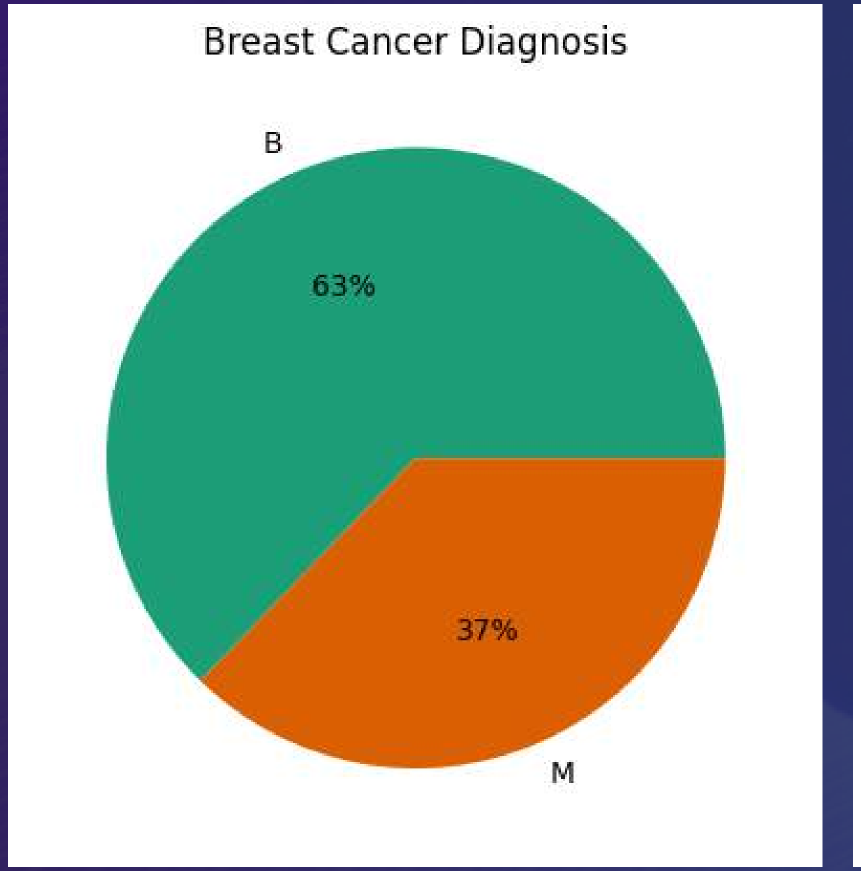


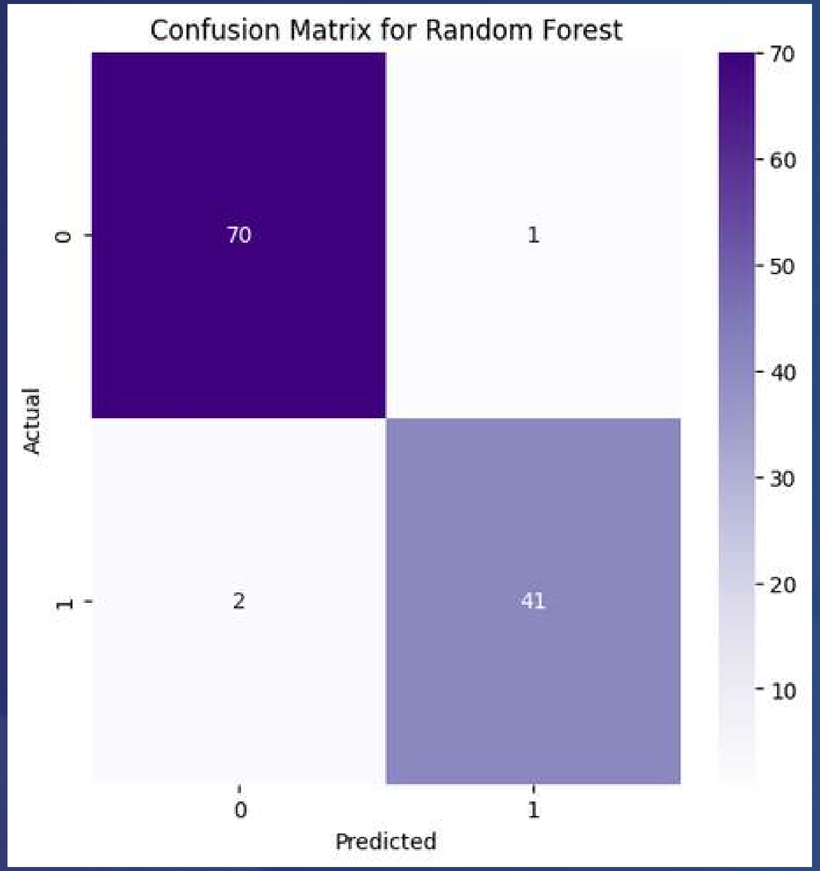
```
              precision    recall  f1-score   support

           0       0.90      0.90      0.90       173
           1       0.55      0.54      0.55        39

    accuracy                           0.83       212
   macro avg       0.72      0.72      0.72       212
weighted avg       0.83      0.83      0.83       212
```

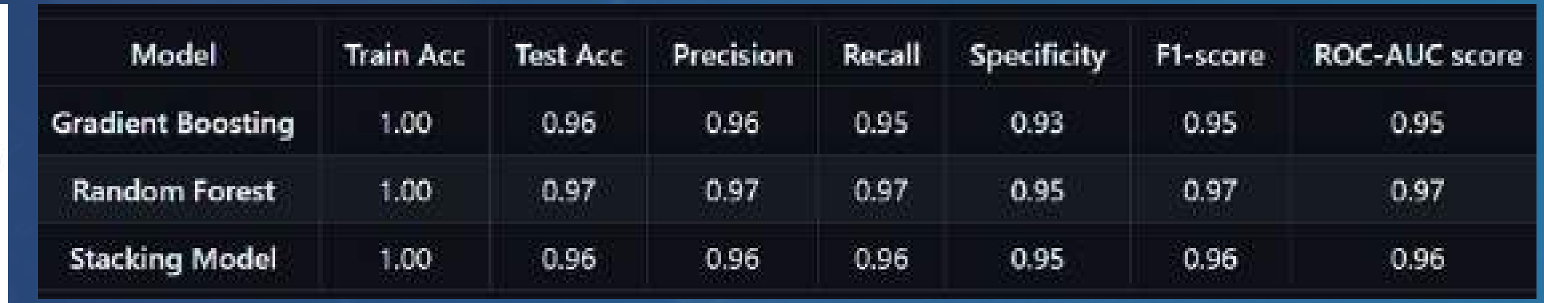
FULL PROJECT HERE

BREAST CANCER PREDICTION

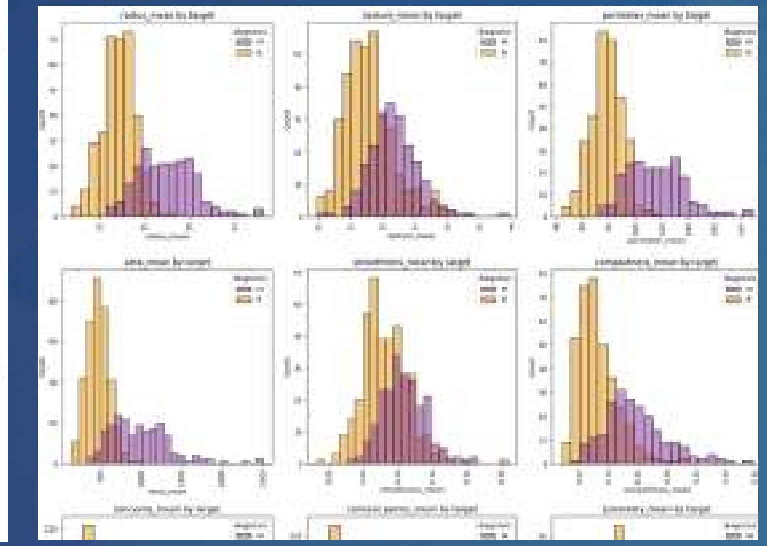
Employed data analysis to identify patterns within the breast cancer dataset. Developed a breast cancer prediction model utilizing the Random Forest algorithm to enhance accuracy and reliability. Two additional models, Gradient Boosting and Stacking, were compared to evaluate performance. Compared against the two models, Random Forest emerged as the top performer with 97% testing accuracy, providing a robust tool for early detection and improved patient outcomes.

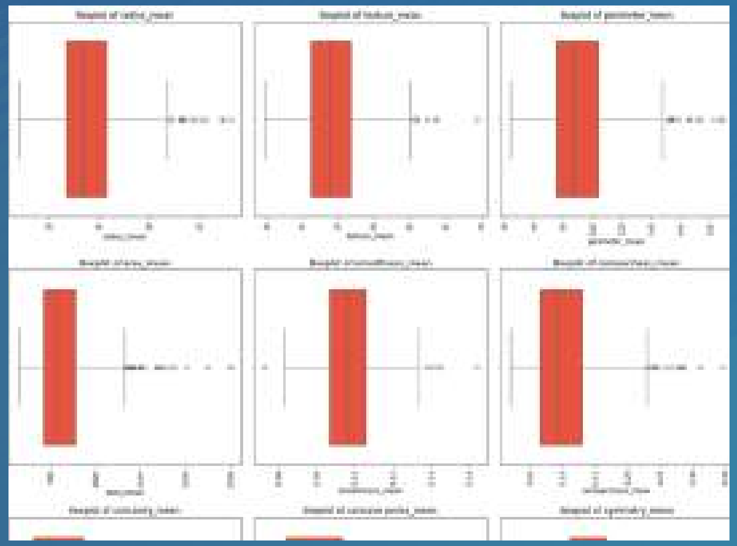






Model	Train Acc	Test Acc	Precision	Recall	Specificity	F1-score	ROC-AUC score
Gradient Boosting	1.00	0.96	0.96	0.95	0.93	0.95	0.95
Random Forest	1.00	0.97	0.97	0.97	0.95	0.97	0.97
Stacking Model	1.00	0.96	0.96	0.96	0.95	0.96	0.96

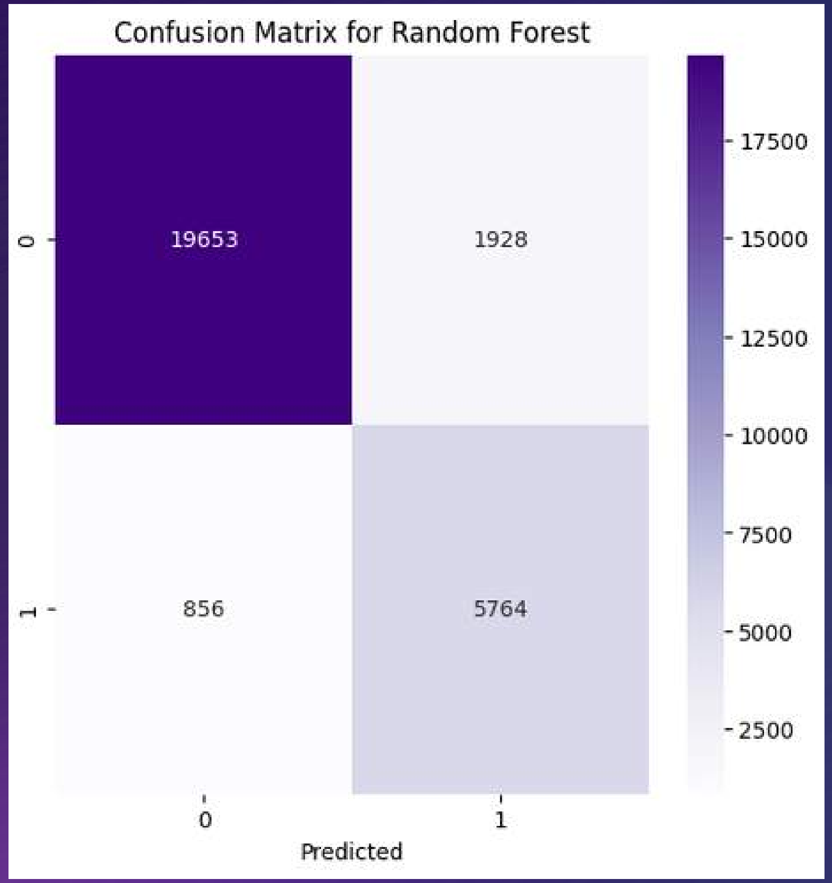


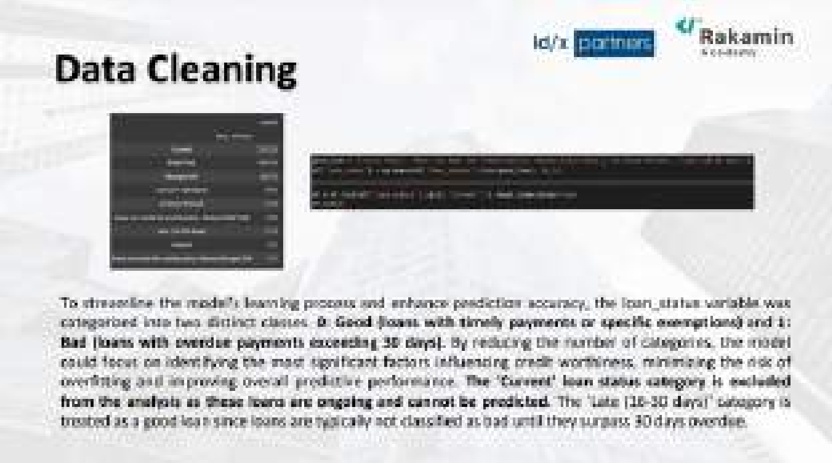


FULL PROJECT HERE

CREDIT RISK PREDICTION

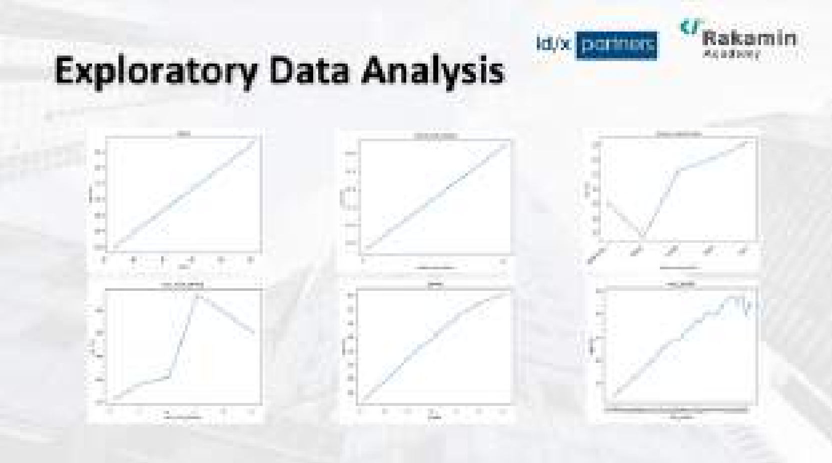
Leveraged Python for exploratory data analysis (EDA) on over 400,000 credit risk data rows spanning 7 years, uncovering actionable insights to enhance company operations. The project focused on analyzing patterns of bad loans and predicting credit risk, with the goal of reducing the instance of bad credit due to a high default rate compared to industry standards. Addressed data imbalance using SMOTE oversampling and developed a Random Forest model that achieved 90% accuracy and an AUC of 89%, enabling more informed, data-driven decisions and improving credit risk management.





Data Cleaning

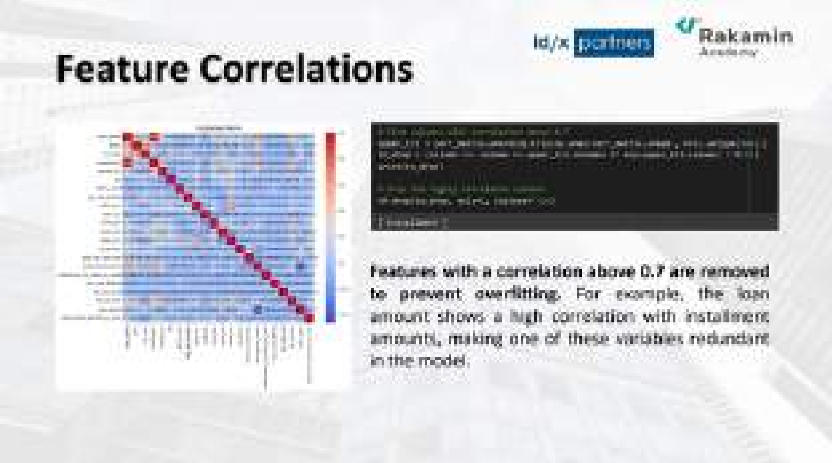
To streamline the model's learning process and enhance prediction accuracy, the loan_status variable was categorized into two distinct classes: **0: Good (loans with timely payments or specific exemptions) and 1: Bad (loans with overdue payments exceeding 30 days).** By reducing the number of categories, the model could focus on identifying the most significant factors influencing credit worthiness, minimizing the risk of overfitting and improving overall predictive performance. **The 'Current' loan status category is excluded from the analysis as these loans are ongoing and cannot be predicted.** The 'Late (16-30 days)' category is treated as a good loan since loans are typically not classified as bad until they surpass 30 days overdue.



Exploratory Data Analysis

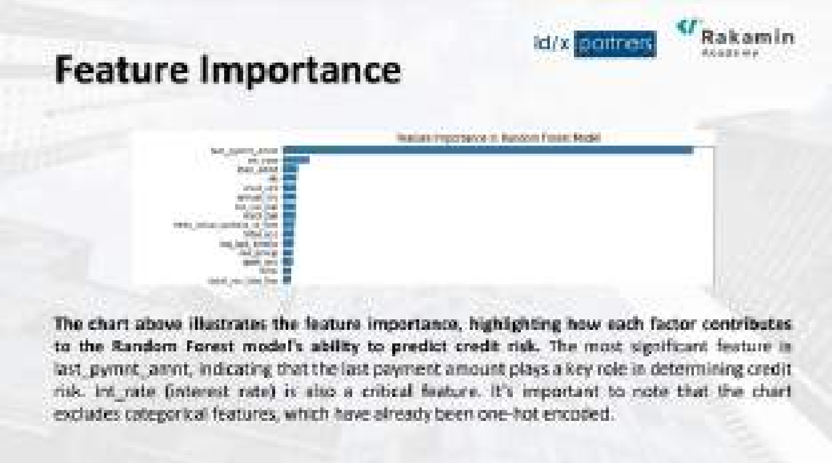
Exploratory Data Analysis

The high default rate of 26% in 2007 can likely be attributed to the onset of the global financial crisis, which started in that year. The housing market collapse, widespread foreclosures, and tightening credit conditions caused significant financial strain, leading to a spike in defaults. **In 2008, the default rate decreased to 20.7%, which, while still high, reflects the ongoing effects of the crisis as governments and financial institutions implemented emergency measures,** including bailouts and loan modifications, to stabilize the economy.



Feature Correlations

Features with a correlation above 0.7 are removed to prevent overfitting. For example, the loan amount shows a high correlation with installment amounts, making one of these variables redundant in the model.



Feature Importance

The chart above illustrates the feature importance, highlighting how each factor contributes to the Random Forest model's ability to predict credit risk. The most significant feature is last_pymnt_amnt, indicating that the last payment amount plays a key role in determining credit risk. Int_rate (interest rate) is also a critical feature. It's important to note that the chart excludes categorical features, which have already been one-hot encoded.

FULL PROJECT HERE

FULL SLIDE