

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.

Scan for a bonus!



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



Machine Learning Specialization



TensorFlow Developer by DeepLearning.AI



TensorFlow: Data and Deployment



TensorFlow: Advanced Techniques



Google Data Analytics Professional Certificate



Data Analysis with Python

CERTIFICATES



Machine Learning Implementation



Machine Learning Operations (MLOps)



Accenture Data Analytics Job Simulation



PwC Switzerland Power BI Job Simulation



ASEAN DSE 2024 Enablement Session



ASEAN DSE 2024 National Finalist

The background features a dark blue gradient with a faint, semi-transparent image of a laptop. In the top-left and bottom-right corners, there are abstract, glossy, pink and purple shapes that resemble liquid or soft clay, with thin, curved lines and small spheres floating around them.

PROJECTS

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

DIABETES IN ASEAN ANALYSIS



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



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.

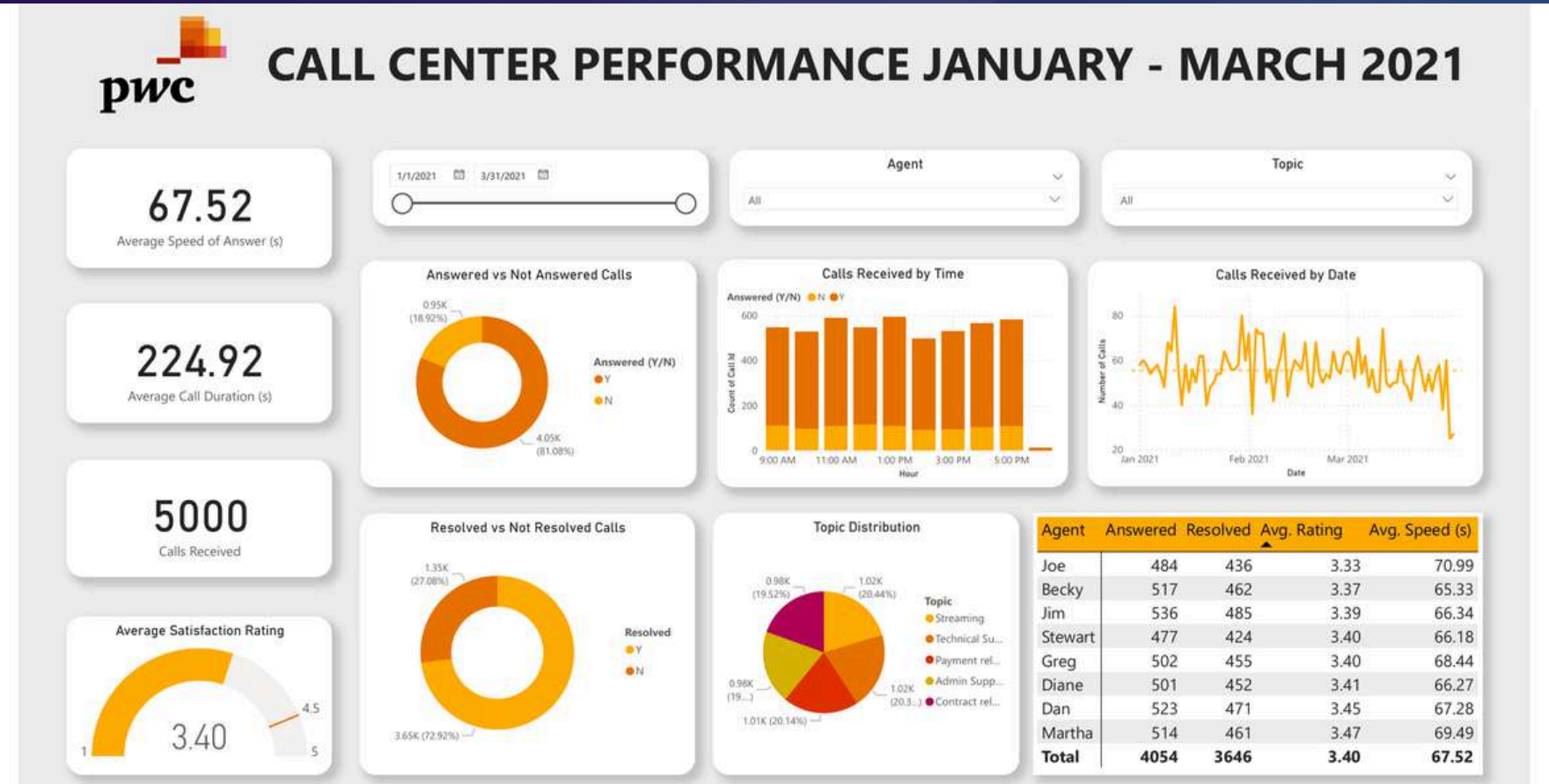
```
CREATE TABLE base_table (  
  SELECT  
  CONCAT(penjualan.id_invoice, penjualan.id_barang) AS id_penjualan,  
  penjualan.id_invoice,  
  penjualan.tanggal,  
  penjualan.id_customer,  
  pelanggan.level,  
  pelanggan.nama,  
  pelanggan.id_cabang,  
  pelanggan.cabang_sales,  
  pelanggan.id_distributor,  
  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_penjual	tanggal	bulan	id_custon	nama	grup	cabang	seid_barang	nama_barj	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	KLORPRO	40	TABLET	SLCYL	47000	1880000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SA	Klinik	Tangerang	BRG0004	TRAMADC	10	TABLET	VNS	24500	245000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SA	Klinik	Tangerang	BRG0005	KLORPRO	10	TABLET	SLCYL	47000	470000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SA	Klinik	Tangerang	BRG0006	KETOCON	10	TABLET	OGB & PH	39000	390000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SA	Klinik	Tangerang	BRG0007	ERGOTAM	10	BOTOL	ETIKAL	64700	647000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SA	Klinik	Tangerang	BRG0008	TETRACYC	10	TABLET	MARCKS	9800	98000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SA	Klinik	Tangerang	BRG0009	AMBROXC	67	BOTOL	VNS	31000	2077000
IN6023BR	2/1/2022	February	CUST5539	KLINIK SA	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	KLORPRO	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 M	Apotek	Bandung	BRG0001	ACYCLOVI	98	DUS	OGB & PH	96000	9408000
IN6028BR	1/30/2022	January	CUST5542	APOTEK M	Apotek	Bandung	BRG0002	ALERGINE	5	DUS	ETIKAL	112000	560000


FULL PROJECT HERE

LOOKER DASHBOARD

CALL CENTER DASHBOARD



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.

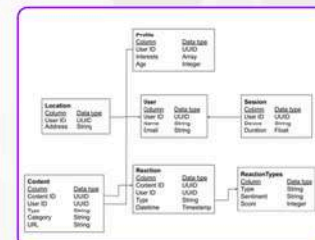


PROJECT & DATA UNDERSTANDING

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 24529 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 1007 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 "actual" content the most.

May 2021 was the month with the most post reactions with a total of 2134 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.

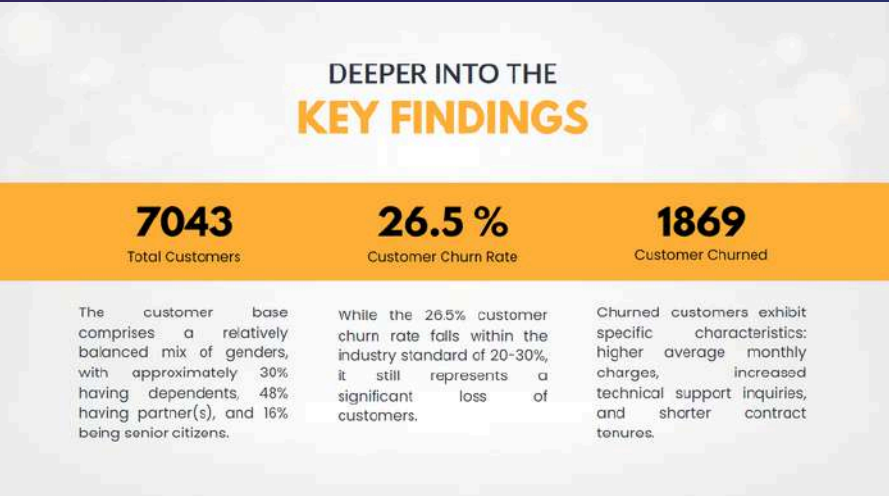
This ad-hoc analysis is insightful, but it's time to take this analysis into large scale production for real-time understanding of the business.

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.

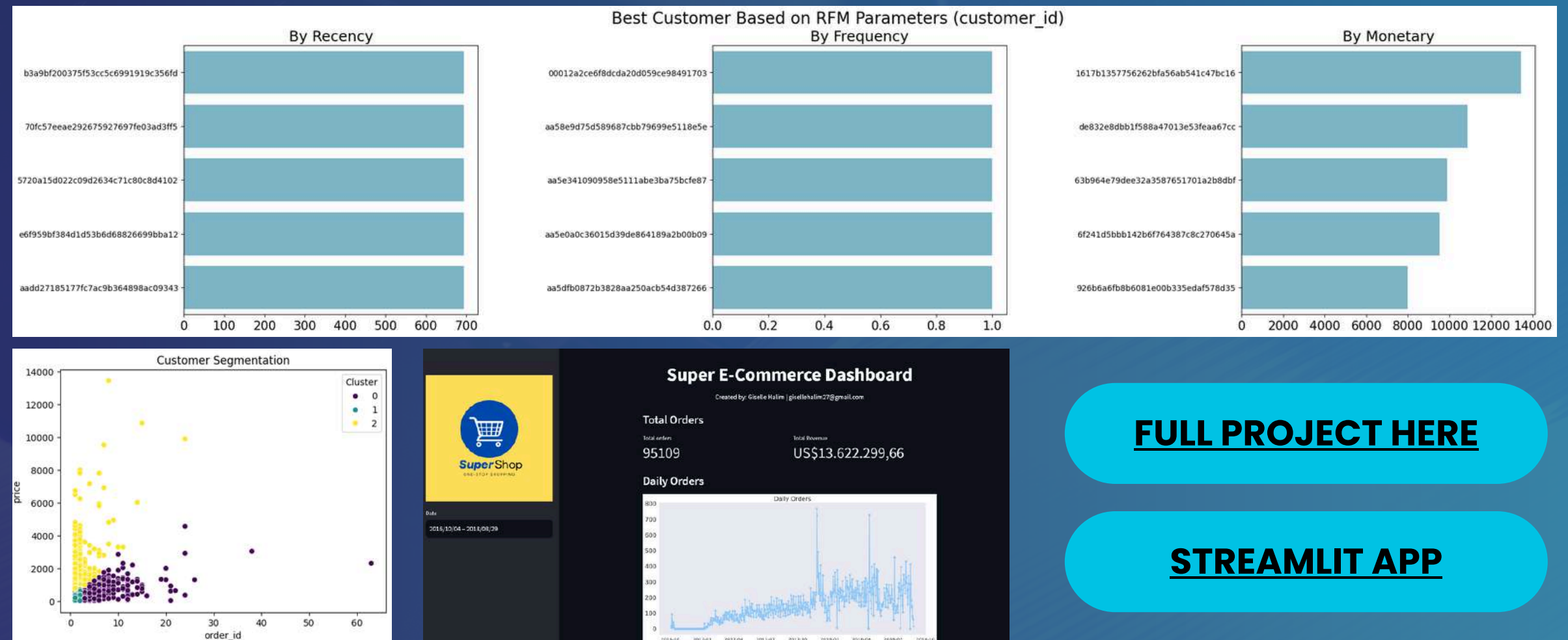


E-COMMERCE SALES ANALYSIS

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?

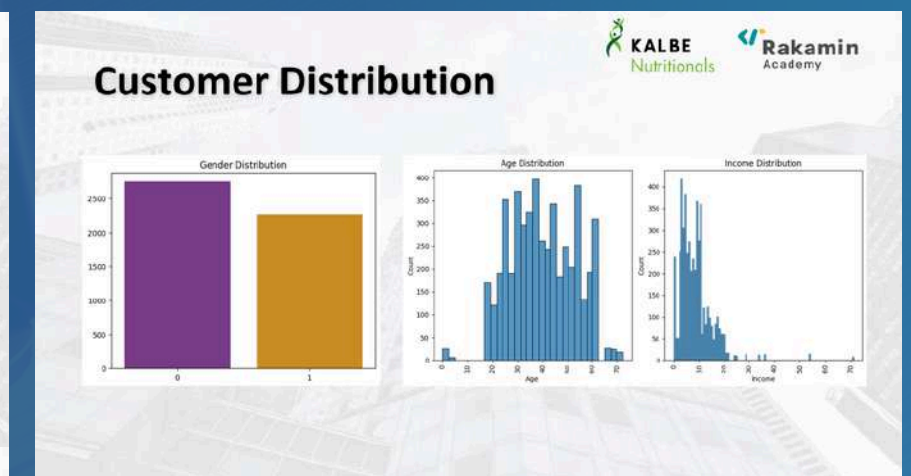
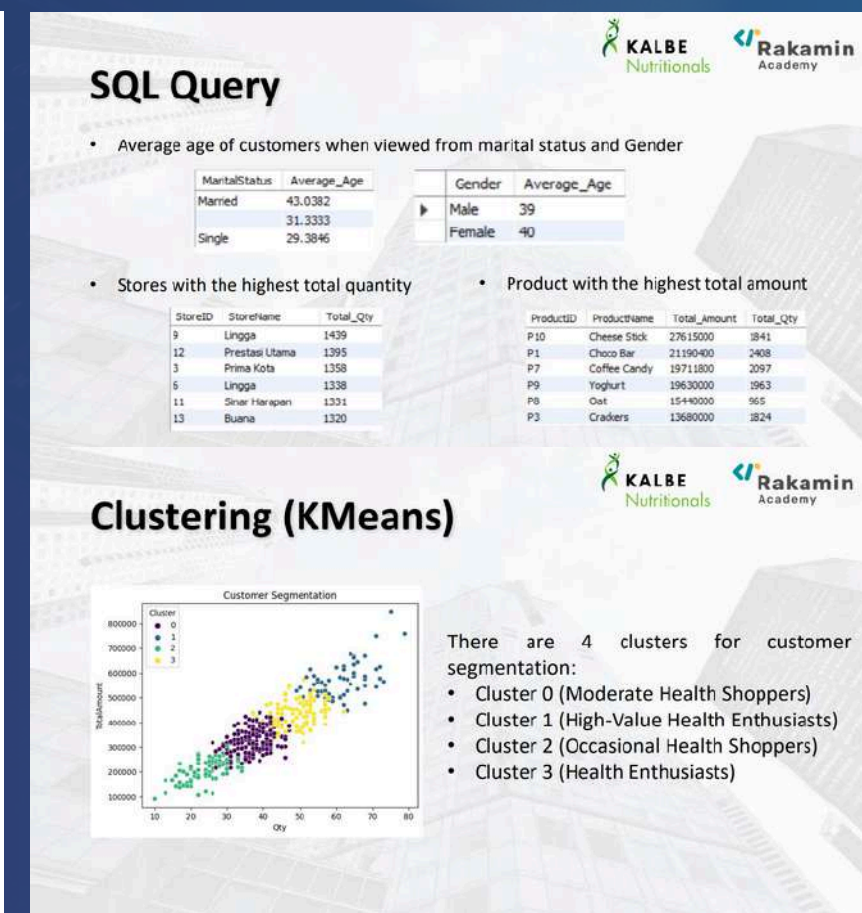
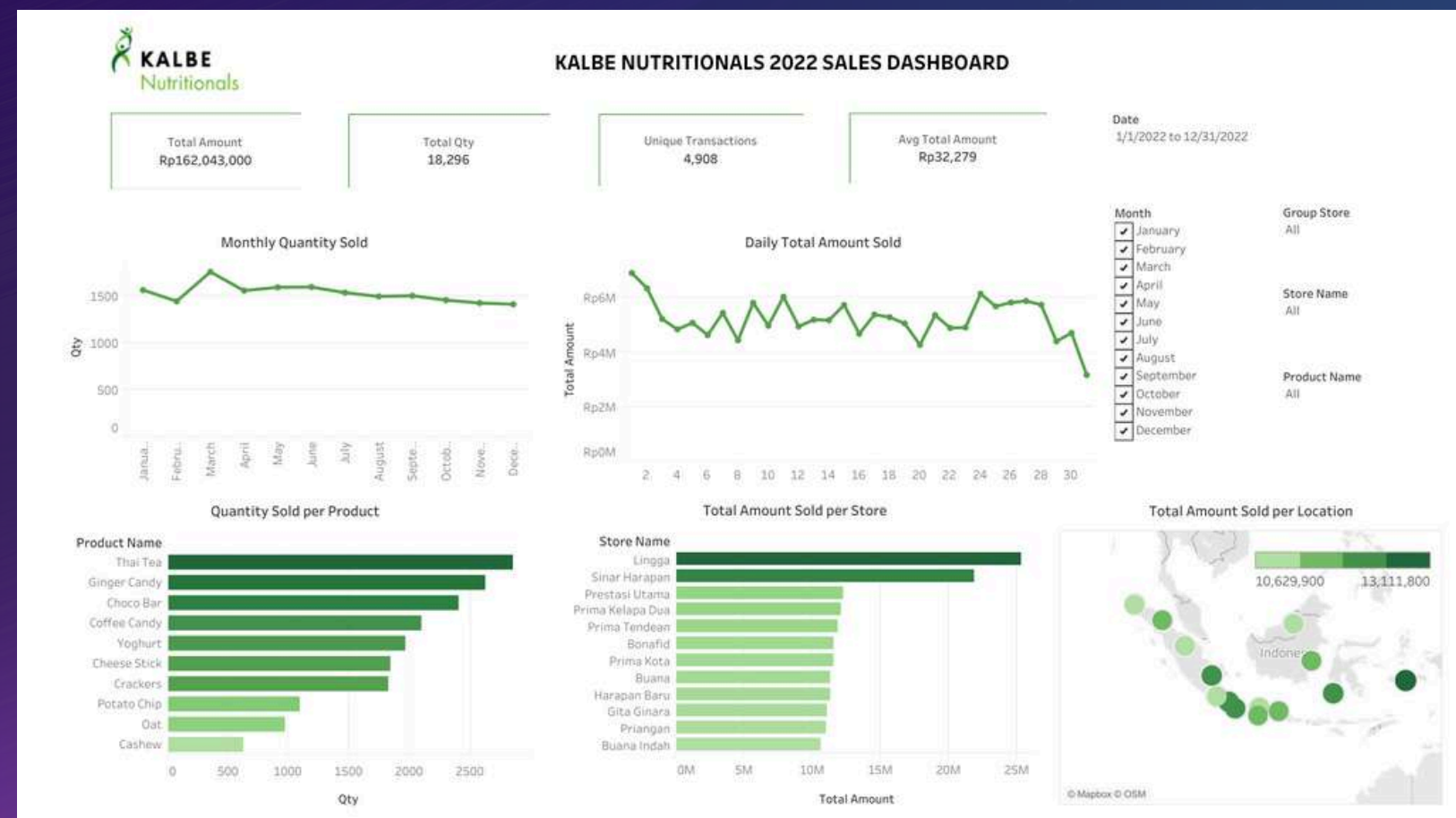


[FULL PROJECT HERE](#)

[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.




[FULL PROJECT HERE](#)

[TABLEAU DASHBOARD](#)

[FULL SLIDE](#)

CERVICAL CANCER RISK IDENTIFICATION WEB-APP

The Cervical Cancer Risk Identification Web App was created for my thesis. Developed using Python, HTML, CSS, and Flask, this web app leverages random forest algorithm to accurately assess cervical cancer risk with machine learning. By analyzing user-provided data on behaviors and medical history, the web app provides valuable insights into an individual's risk level, promoting awareness and empowering users to take proactive steps for their health. With a remarkable 96.5% accuracy, this innovative tool serves as a valuable resource for individuals seeking information about cervical cancer prevention.

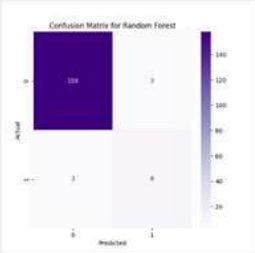


Introduction

Cervical cancer is one of the most common types of cancer affecting women in Indonesia. The good news is that cervical cancer can be prevented and treated effectively if detected early. Cervix-Intel is here to help you understand your risk of cervical cancer, factors that may increase the risk, and preventive measures you can take to protect yourself.

Cervix-Intel Technology

Cervix-Intel uses machine learning technology, specifically the random forest algorithm, to identify the risk of cervical cancer. The model achieved an accuracy of **96.5%** during testing, 0.73 for recall, precision, and f1-score. While the model's AUC-ROC score reached **0.85**. The confusion matrix is shown in the figure on the side.



Cervical Cancer Risk Assessment

Fill out the questionnaire below to determine your risk of cervical cancer. This assessment does not provide a definitive diagnosis and can only be used as preliminary information.

Age:

Sexual Intercourse and Pregnancy

Number of sexual partners:

First sexual intercourse:

Number of pregnancies:

Smoking Habits

Do you smoke?

How many years have you been smoking?

How many packs a year?

Contraceptive Usage

Are you using hormonal contraceptives?

How many years have you been using hormonal contraceptives?

Are you using IUD?

How many years have you been using IUD?

History of Reproductive System Disease

How many STD diagnoses did you receive?

Have you ever had condylomatosis or genital warts?

Have you ever had syphilis?

Have you ever had pelvic inflammatory disease?

Have you ever had genital herpes?

Do you have HIV?

Have you ever been diagnosed with cervical dysplasia?

Have you ever been diagnosed with HPV?

Reproductive Organ Health Check Results

How are your VIA test results?

How are your schiller test results?

How are your pap smear/cervical cytology test results?

Your Cervical Cancer Risk Assessment Results

Low Risk

77.0%
Low Risk

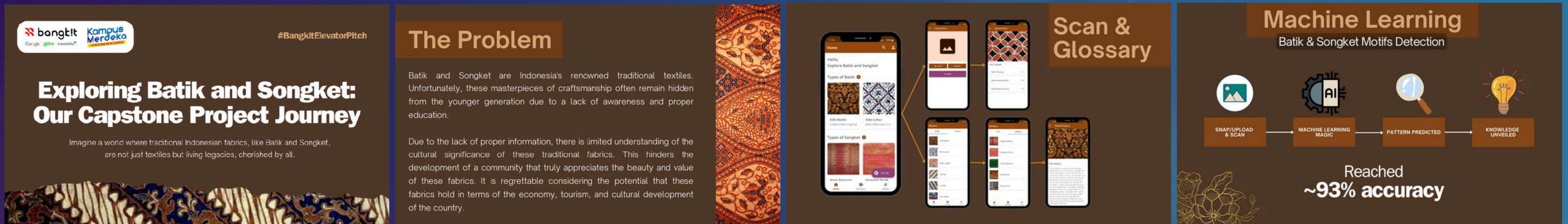
23.0%
High Risk

What does it mean?

Based on your risk assessment results, you have a low risk of cervical cancer. This means that your chances of developing cervical cancer at this time are low. However, it is important to remember that this result does not guarantee you are completely risk-free. Pap smears are still recommended regularly to detect abnormal cell changes that may lead to cervical cancer at an early stage. In addition, maintaining a healthy lifestyle with a balanced diet, regular physical activity, and avoiding smoking, can help keep your reproductive organs healthy and lower your risk of cervical cancer in the future. You can seek more information about cervical cancer and its prevention from reliable sources such as health organizations or official websites. Having a low risk is good news. Stay healthy and enjoy a healthy lifestyle.

PUSAKA NUSANTARA

This project, undertaken as part of an independent study at Bangkit Academy, focused on developing an Android app capable of accurately identifying traditional Indonesian fabrics through photo, specifically batik and songket motifs. I took part as the machine learning team in this project. Utilizing a Convolutional Neural Network (CNN) architecture with TensorFlow Keras and leveraging transfer learning from InceptionV3, the model achieved an impressive 93% accuracy in differentiating between the patterns. To ensure seamless integration into the Android app, the trained model was converted to the TensorFlow.js format.



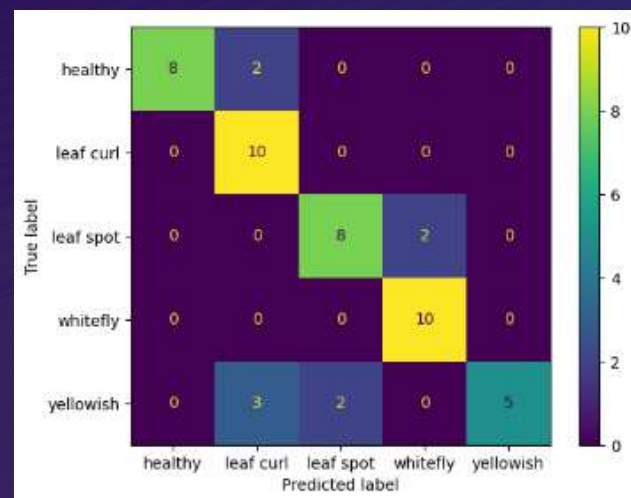
```
Epoch 1/5
38/38 [=====] - 6s 75ms/step - loss: 1.3741 - acc: 0.6870 - val_loss: 0.6830 - val_acc: 0.7000
Epoch 2/5
38/38 [=====] - 2s 52ms/step - loss: 0.1363 - acc: 0.9629 - val_loss: 0.3719 - val_acc: 0.8750
Epoch 3/5
38/38 [=====] - 2s 58ms/step - loss: 0.0140 - acc: 0.9947 - val_loss: 0.3838 - val_acc: 0.9250
Epoch 4/5
38/38 [=====] - 2s 47ms/step - loss: 0.0124 - acc: 0.9947 - val_loss: 0.2867 - val_acc: 0.9250
Epoch 5/5
38/38 [=====] - 2s 47ms/step - loss: 2.7509e-04 - acc: 1.0000 - val_loss: 0.2632 - val_acc: 0.9250
```

[FULL PROJECT HERE](#)

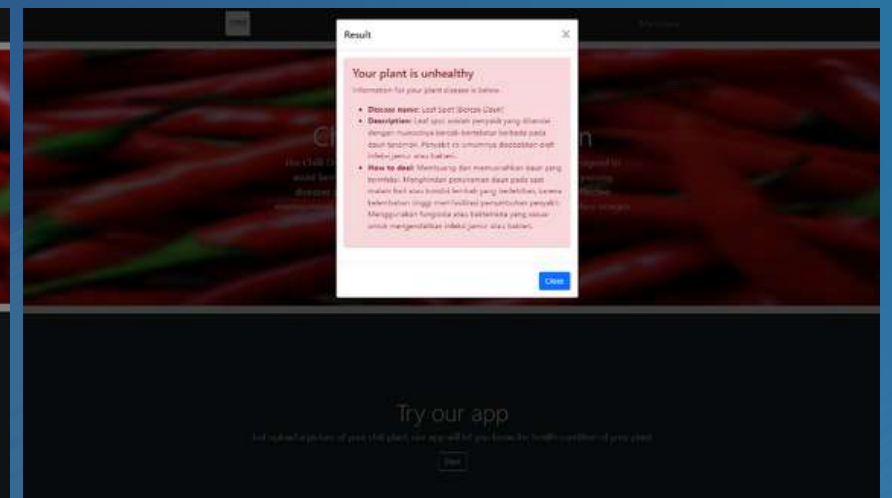
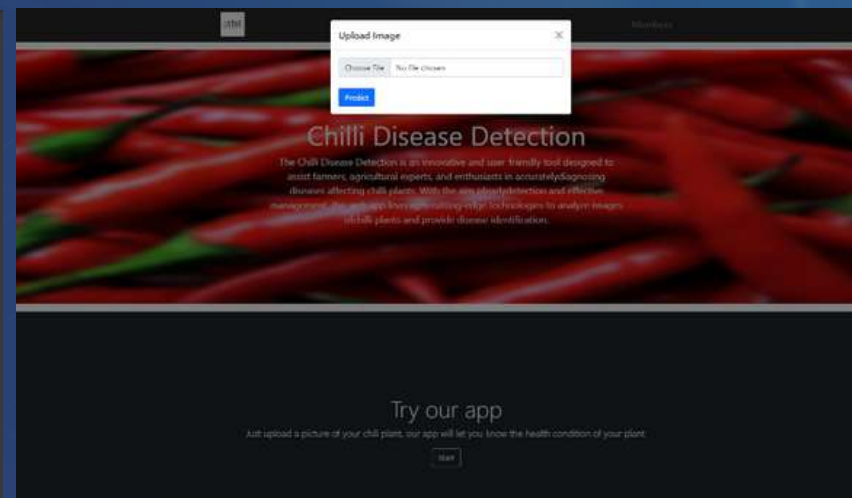
[ELEVATOR PITCH](#)

CHILI DISEASE DETECTION

As part of an independent study at Orbit Future Academy, I assumed the role of a machine learning engineer in the capstone project to develop a web application capable of accurately identifying chili plant diseases through image analysis. Utilizing the TensorFlow Keras framework and leveraging transfer learning with MobileNet, I constructed a robust machine learning model that achieved an impressive 82% test accuracy. This model effectively differentiates healthy chili leaves from four common diseases, providing insights for farmers and agricultural professionals.



	precision	recall	f1-score	support
0	1.00	0.80	0.89	10
1	0.67	1.00	0.80	10
2	0.80	0.80	0.80	10
3	0.83	1.00	0.91	10
4	1.00	0.50	0.67	10
accuracy			0.82	50
macro avg	0.86	0.82	0.81	50
weighted avg	0.86	0.82	0.81	50



[FULL PROJECT HERE](#)

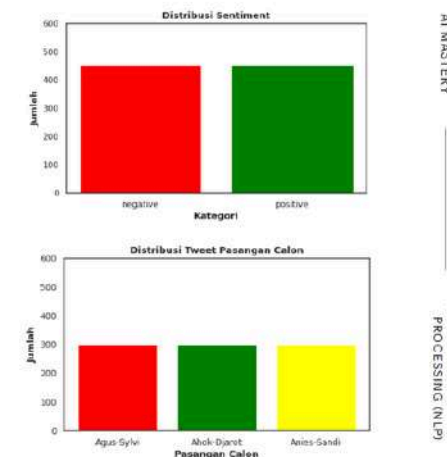
2017 JAKARTA LOCAL ELECTION SENTIMENT ANALYSIS

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.



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.

AI MASTERY

NATURAL LANGUAGE PROCESSING (NLP)

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.

NATURAL LANGUAGE PROCESSING (NLP)

AI MASTERY

Naive Bayes Performance (BernoulliNB)

```
#predicted vs actual label
prediksi_benar = (model_pred == y_test).sum()
prediksi_salah = (model_pred != y_test).sum()

print('Jumlah prediksi benar:', prediksi_benar)
print('Jumlah prediksi salah:', prediksi_salah)

accuracy = prediksi_benar / (prediksi_benar + prediksi_salah)*100
print('Akurasi pengujian:', accuracy, '%')
```

```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, model_pred)
print('Confusion matrix:\n', cm)

Confusion matrix:
[[78 17]
 [ 5 88]]
```

```
# Cross Validation
from sklearn.model_selection import StratifiedShuffleSplit
from sklearn.metrics import accuracy_score

cv = StratifiedShuffleSplit(n_splits=5, test_size=0.2, random_state=0)

accuracy = cross_val_score(model, X_test_features, y, cv=cv, scoring='accuracy')
avg_accuracy = np.mean(accuracy)

print('Akurasi setiap split:', avg_accuracy, '%')
print('Data-rata akurasi data cross validation:', avg_accuracy)
```

Classification report:				
	precision	recall	F1-score	support
-1	0.54	0.82	0.68	95
1	0.82	0.94	0.88	85
accuracy			0.88	180
macro avg	0.68	0.88	0.88	180
weighted avg	0.80	0.88	0.88	180

[FULL PROJECT HERE](#)

[FULL SLIDE](#)



BOOK RECOMMENDATION SYSTEM

Employs both content-based and collaborative filtering techniques. Content-based filtering analyzes book features and user preferences to suggest similar titles with TF-IDF vectorization, the system will recommend several books that have the same author or have similar title. Collaborative filtering leverages user behavior and ratings to recommend books that others with similar tastes have enjoyed. TensorFlow Keras were employed for collaborative filtering, further refining the recommendation process. This dual approach ensures personalized and accurate book recommendations.

	litton	modiano	nic	greta	lafferty	paz	lindskold	robie	pacotti	gwendolen	...	rysard	praagh
title													
White Fang	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
One Man's Poison (One Man's Poison)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
Fabulous Nobodies: A Novel About a Girl Who's in Love With Her Clothes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
McNally's Secret (Archy McNally Novels (Paperback))	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
My Left Foot	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

```
# Get user input for the book name
search_name = input("Enter the book title you want to search: ")

# Filter the DataFrame based on the input name
rec = book_recommendations(search_name)
# Misalnya: The Blessing Stone

# Display the results
rec
```

	title	author
0	Virgins of Paradise	Barbara Wood
1	Perfect Harmony	Barbara Wood
2	Bajo El Sol de Kenia	Barbara Wood
3	Dreaming	Barbara Wood
4	Dreaming	Jill Barnett

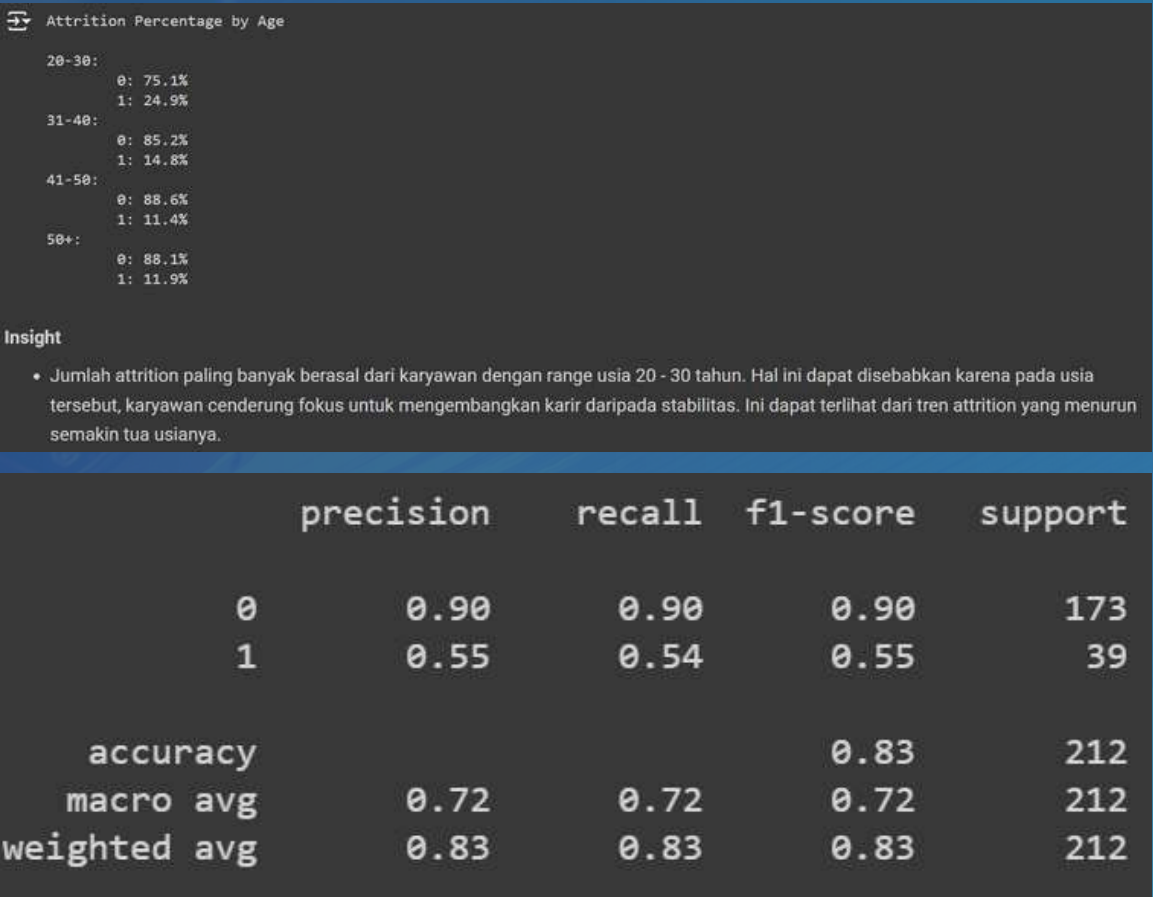
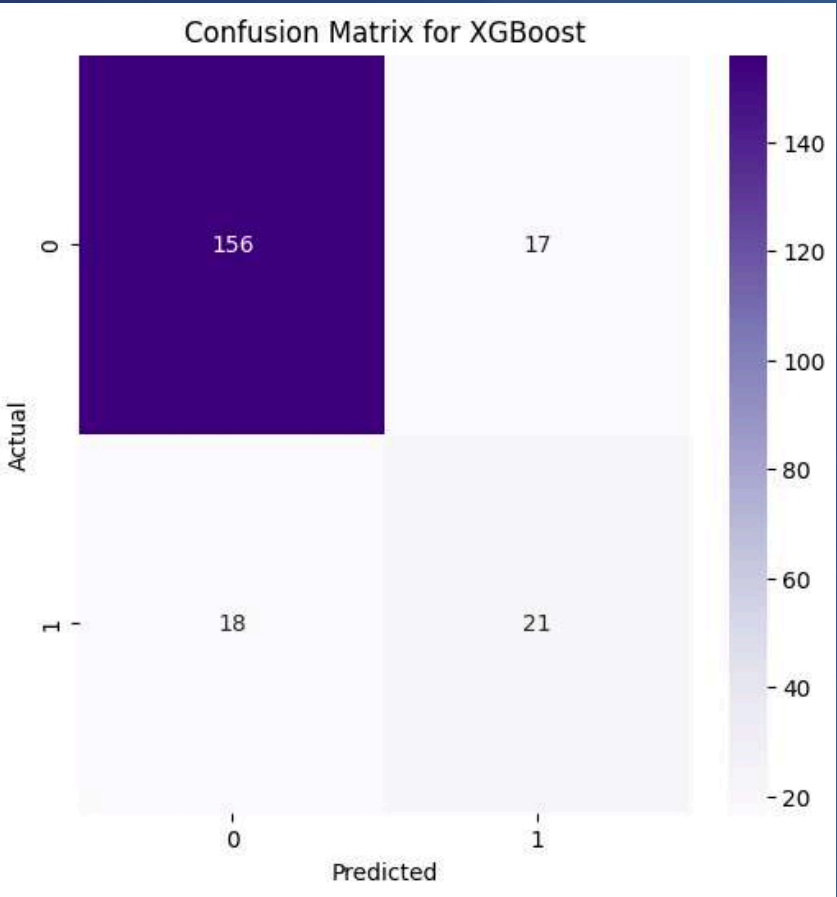
```
458/458 [=====] - 1s 1ms/step
Showing recommendations for users: 432
=====
Book with high ratings from user
-----
Walk Two Moons : Sharon Creech
Wuthering Heights (Wordsworth Classics) : Emily Bronte
This Day All Gods Die: The Gap into Ruin (Gap Series/Stephen R. Donaldson) : Stephen R. Donaldson
Visible Heart (Silhouette Romances #275) : Dixie Browning
Heaven's Price : Sandra Brown
-----
Top 10 Book Recommendation
-----
The Body Farm : Patricia Daniels Cornwell
The Secret (Animorphs, No 9) : Katherine Applegate
Heir to the Shadows (The Black Jewels Trilogy, Book 2) : Anne Bishop
Si c'est un homme : Primo Levi
```

```
Epoch 7/10
6266/6266 [=====] - 72s 11ms/step - loss: 0.5005 - root_mean_squared_error: 0.1530 - val_loss: 0.5225 - val_root_mean_squared_error: 0.1772
Epoch 8/10
6266/6266 [=====] - 71s 11ms/step - loss: 0.4988 - root_mean_squared_error: 0.1513 - val_loss: 0.5208 - val_root_mean_squared_error: 0.1753
Epoch 9/10
6266/6266 [=====] - 72s 11ms/step - loss: 0.4972 - root_mean_squared_error: 0.1496 - val_loss: 0.5205 - val_root_mean_squared_error: 0.1753
Epoch 10/10
6266/6266 [=====] - 71s 11ms/step - loss: 0.4955 - root_mean_squared_error: 0.1477 - val_loss: 0.5204 - val_root_mean_squared_error: 0.1753
```

[FULL PROJECT HERE](#)

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.

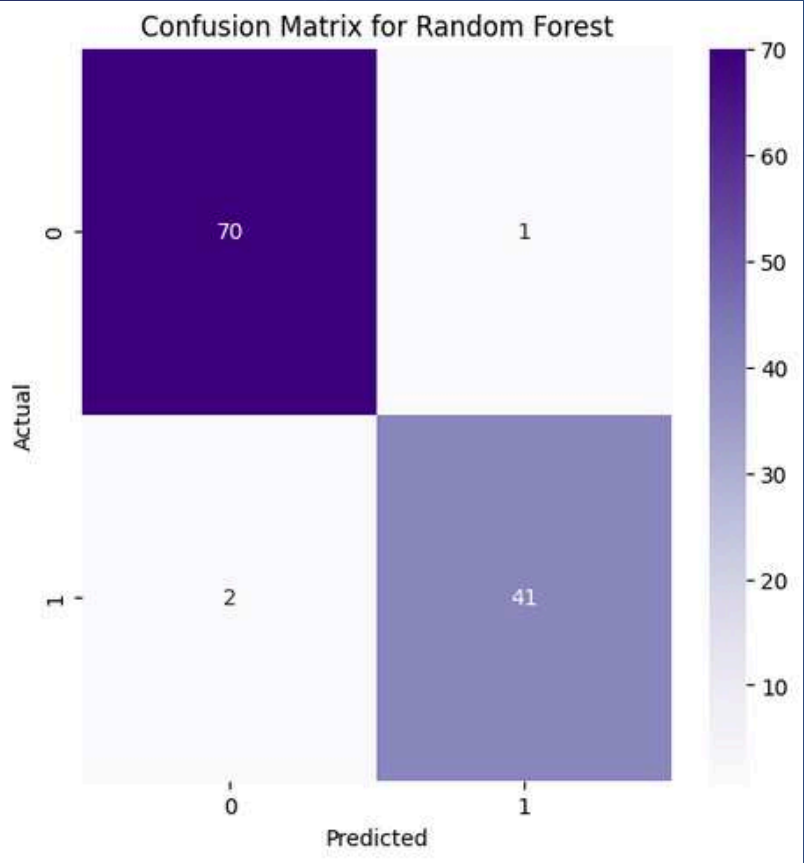
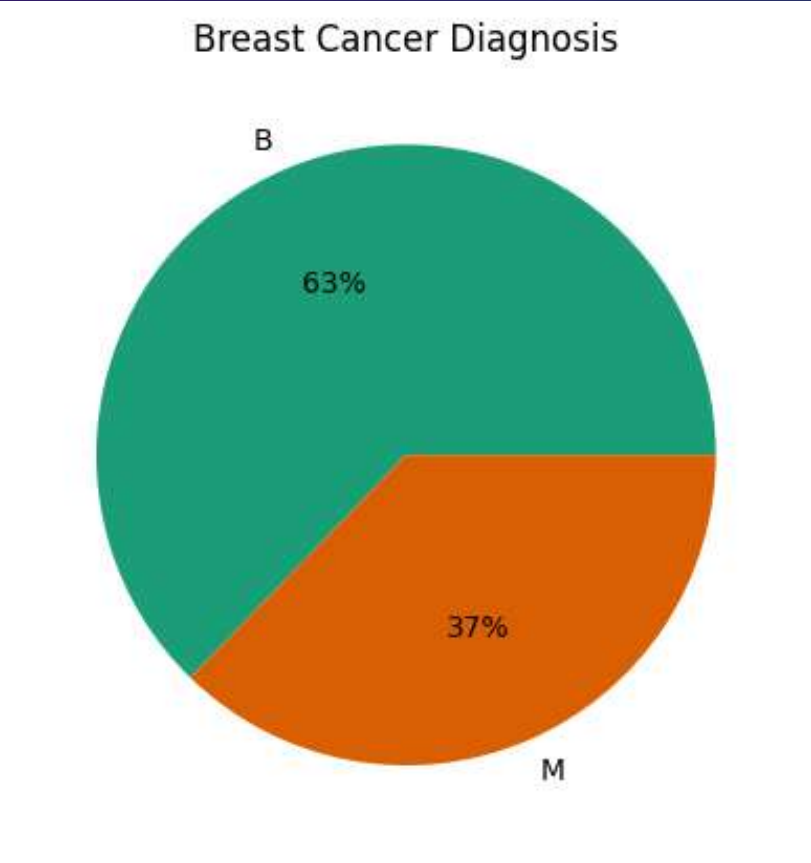


[FULL PROJECT HERE](#)

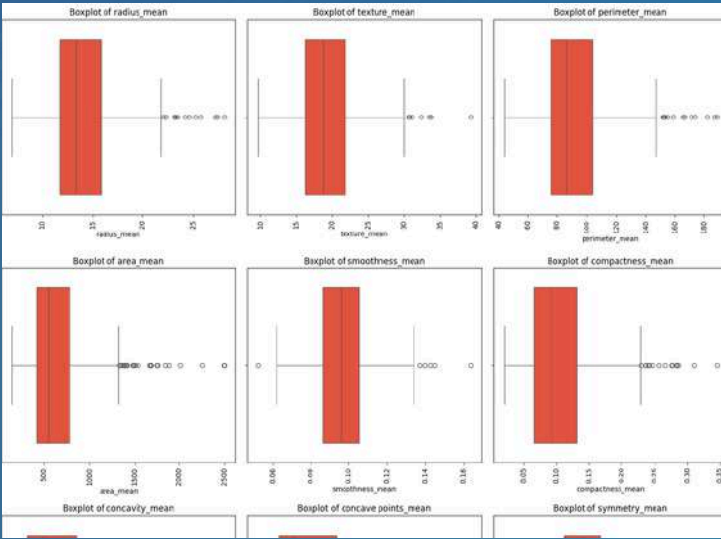
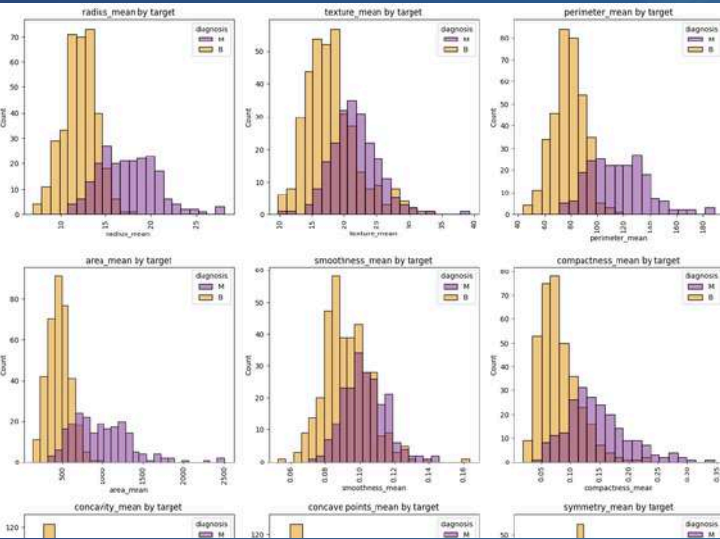


BREAST CANCER PREDICTION

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.



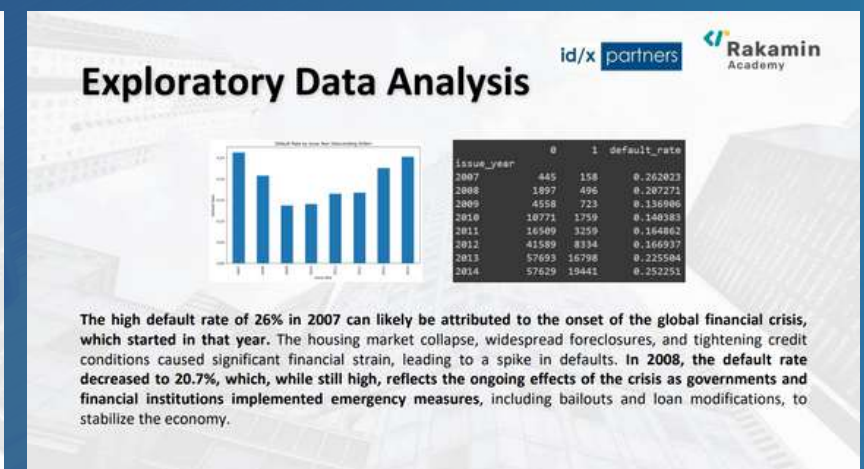
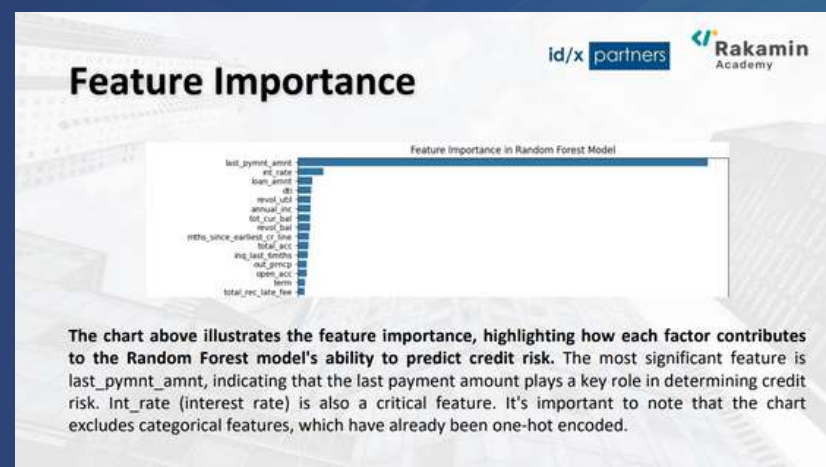
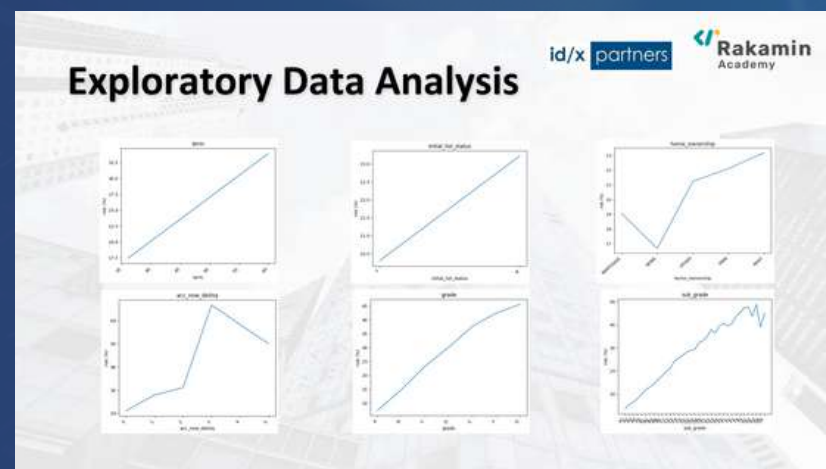
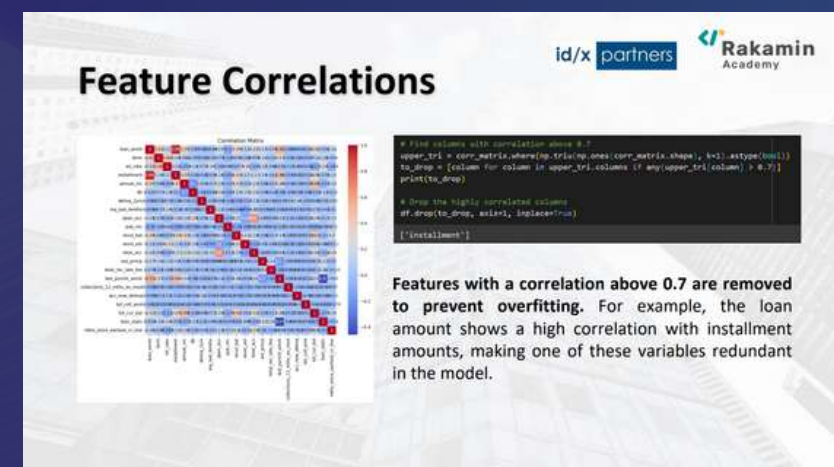
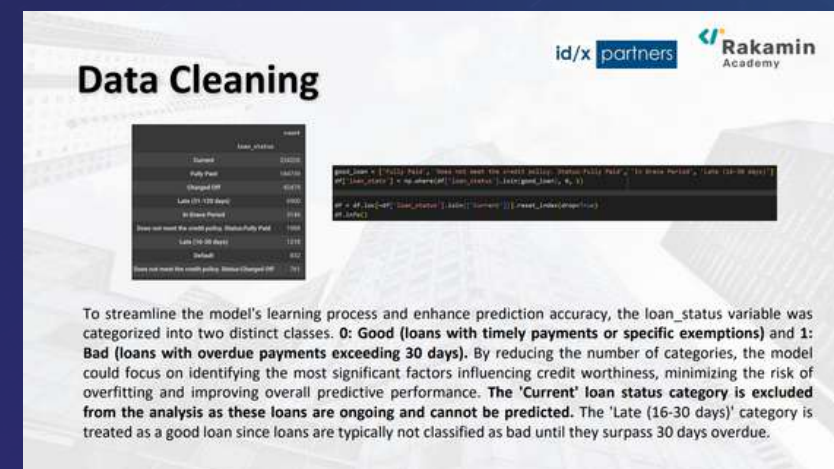
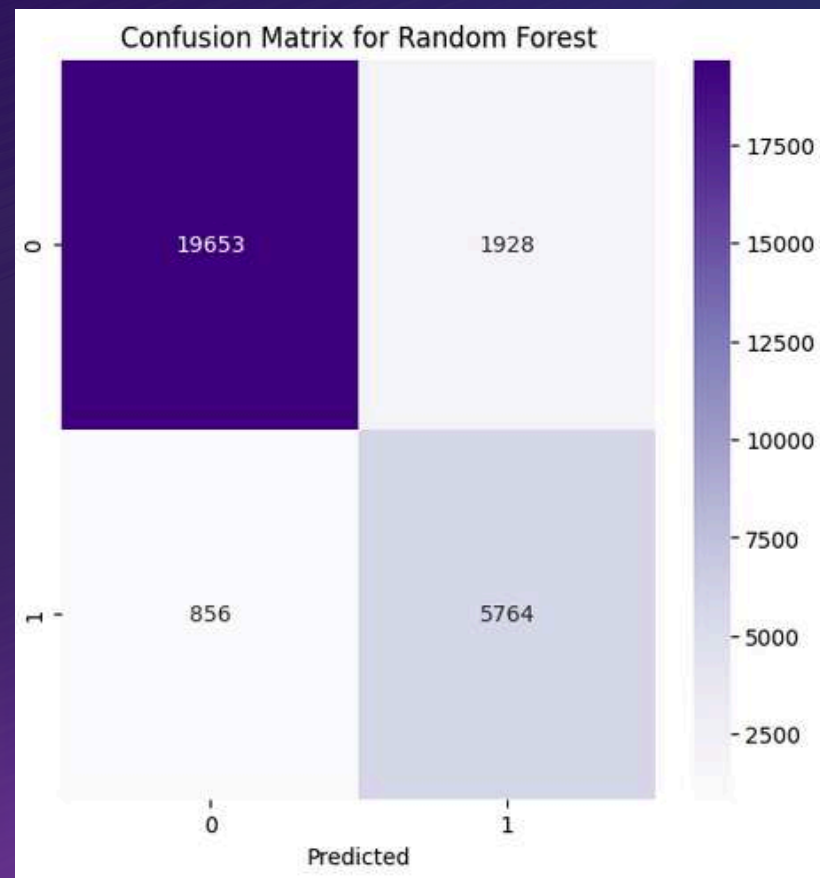
Model	Train Acc	Test Acc	Precision	Recall	Specificity	F1-score	ROC-AUC score
Gradient Boosting	1.00	0.96	0.96	0.95	0.93	0.95	0.95
Random Forest	1.00	0.97	0.97	0.97	0.95	0.97	0.97
Stacking Model	1.00	0.96	0.96	0.96	0.95	0.96	0.96



[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.

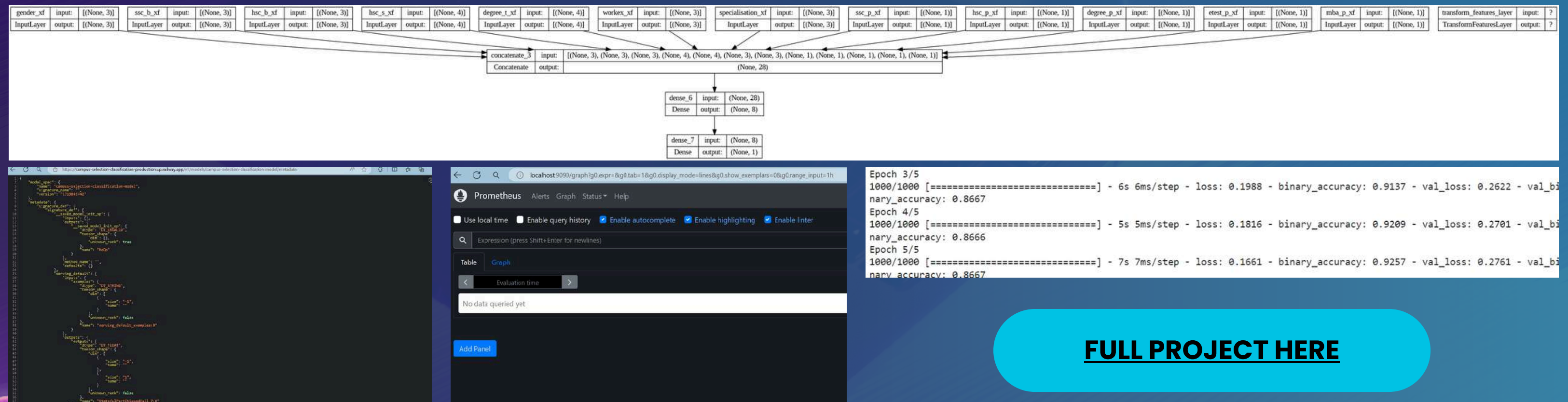


[FULL PROJECT HERE](#)

[FULL SLIDE](#)

CAMPUS SELECTION PIPELINE

This project developed a machine learning pipeline to automate the campus selection process based on education history and work experience. By leveraging TensorFlow Keras and TFX components, the pipeline ensured consistent data preprocessing and training, leading to a highly accurate model. The resulting model achieved an impressive 86.7% accuracy in predicting candidate suitability. To facilitate deployment and monitoring, the model was integrated with the Railway platform and monitored using Prometheus via Docker, ensuring seamless operation and continuous evaluation of performance.





THANK YOU!

GET IN TOUCH WITH ME!

✉ gisellehalim27@gmail.com

 [LinkedIn: Giselle Halim](#)  [GitHub: gisellehalim](#)