EMAIL CLASSIFICATION

NATURAL LANGUAGE PROCESSING

STUDI INDEPENDEN BERSERTIFIKAT - BISA AI ACADEMY

ANDREAN YONATHAN

Kampus Merdeka INDONESIA JAYA

Bisa.ai PT. Bisa Artifisial Indonesia

About Me

Hi, I am Andrean Yonathan you can call me Andrean or Andre. I am a third-year mathematics student at Diponegoro University. I have an interest in technology, especially in artificial intelligence and data science. I have high curiosity, like to learn new things, great in teamwork, and has a sense of responsibility and communication skill.

TRAINING

DQLab, 2021

Learn about data science using Python and SQL

Independent Study at Progate, 2021

Learn about web development using HTML, CSS, JavaScript, and Node.js

Independent Study at Bisa AI Academy, 2022

Learn about artificial intelligence such as data science, natural language processing, machine learning, and image processing

Outline

- Task Overview
- Dataset & Library
- Data Understanding
- Text Pre- processing
- Feature Extraction
- Modelling
- Evaluation

TASK OVERVIEW

In this task, email classification will be carried out using several machine learning algorithms, namely Random Forest Classifier and Logistic Regression. The tool that used in this task is Google Colaboratory. The process starts with getting the data to be used, understanding the data, processing the data, entering the feature extraction process, and then building the model. For evaluation, the metrics to be used are accuracy score, classification report, and confusion matrix.

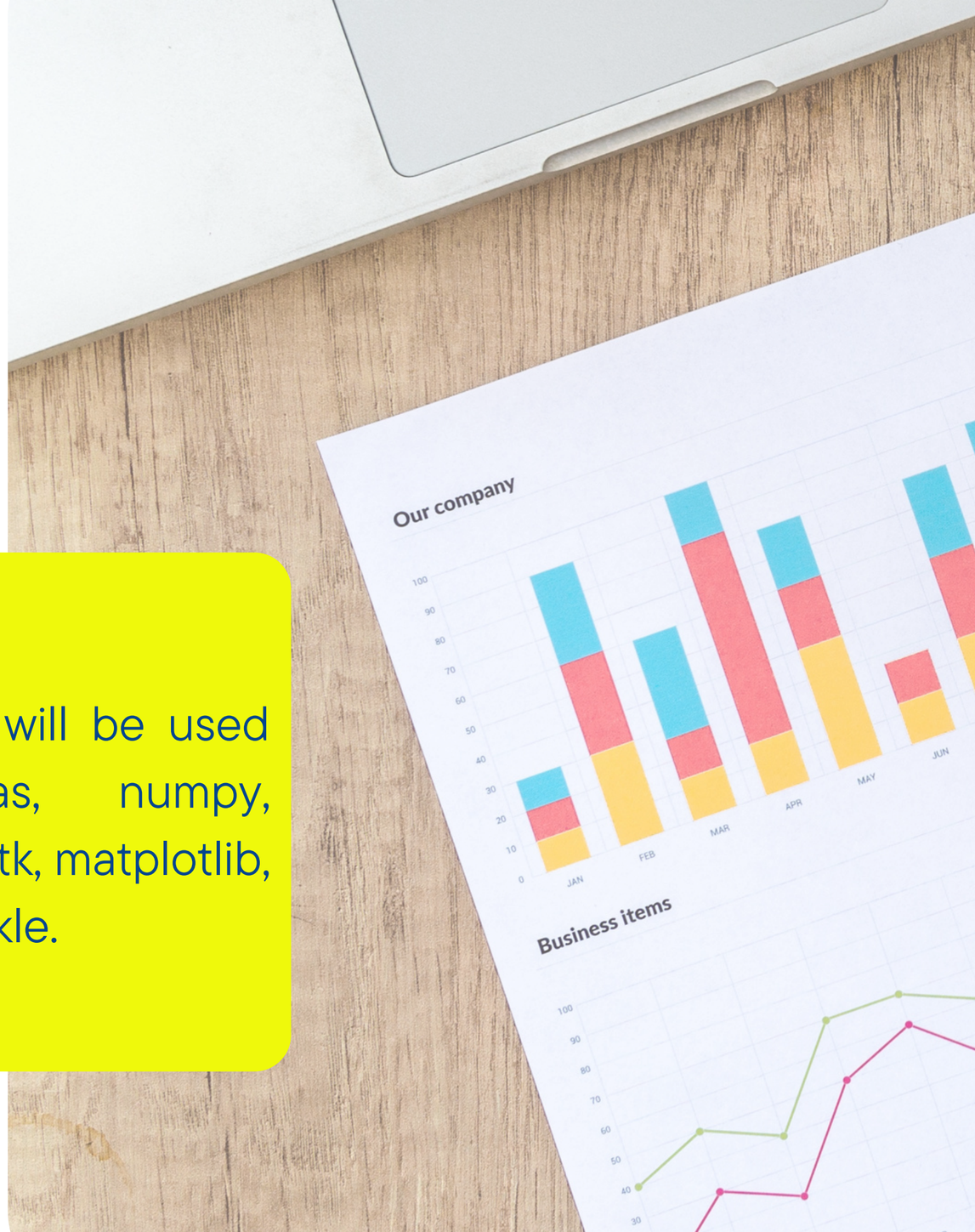
DATASET & LIBRARY

Dataset

The Dataset was obtained from Kaggle open dataset and contains three columns, `S. No.`, `Message_body`, and `Label`

Library

The library that will be used include pandas, numpy, scikit-learn, re, nltk, matplotlib, seaborn, and pickle.



DATASET & LIBRARY

	S. No.	Message_body	Label
0	1	Rofl. Its true to its name	Non-Spam
1	2	The guy did some bitching but I acted like i'd...	Non-Spam
2	3	Pity, * was in mood for that. So...any other s...	Non-Spam
3	4	Will ü b going to esplanade fr home?	Non-Spam
4	5	This is the 2nd time we have tried 2 contact u...	Spam

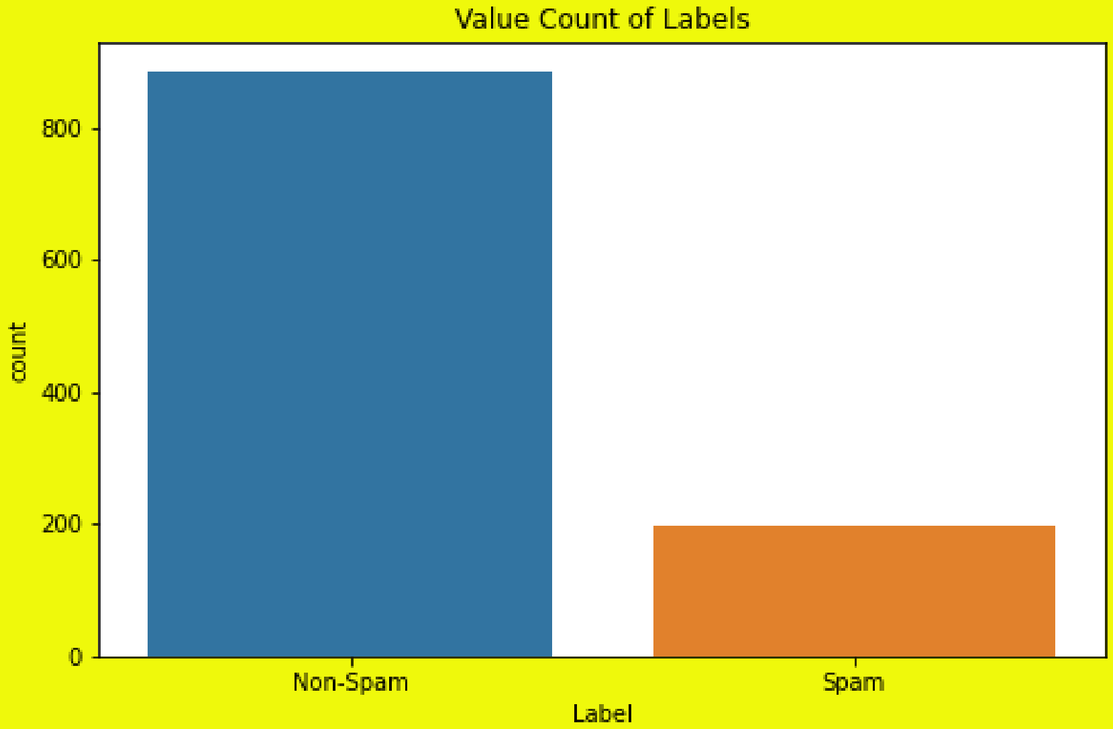
```
# library for data analysist
import pandas as pd
import numpy as np

# library for visualization
import matplotlib.pyplot as plt
import seaborn as sns

# library for text pre-processing
import re
import nltk
nltk.download('stopwords')
from sklearn.preprocessing import LabelEncoder
from nltk.corpus import stopwords
from nltk.stem import *

# library for features extraction
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import CountVectorizer

# library for build and evaluate model
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
from sklearn.metrics import confusion_matrix

# library for save model
import pickle
```

DATA UNDERSTANDING

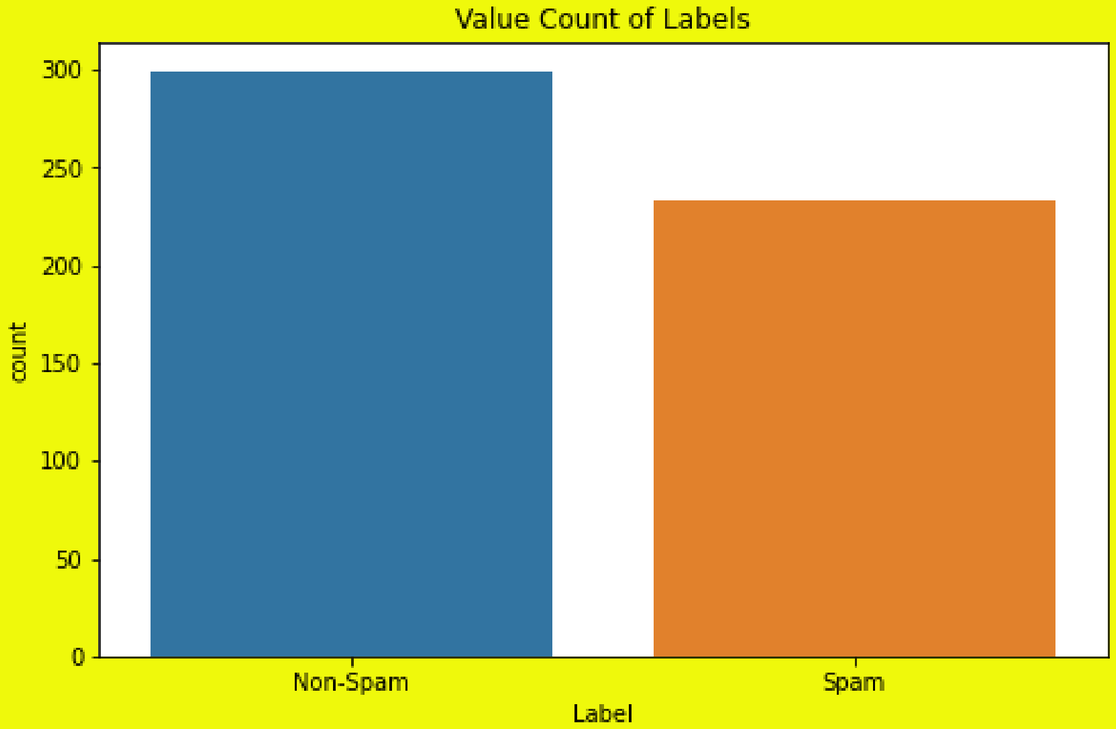
This stage is the process to understand the data and identify problems in the data such as whether there are missing values in the data, whether there are outliers, and so on. From this process, some information was obtained:

1. The data has 1082 records and 3 columns.
2. Column 'S. No.' is an unnecessary column and will be removed.
3. Column 'Label' has object data type and label encoding process will be performed.
4. Data has no missing value.
5. The data is an imbalance dataset and an undersampling process will be carried out.

DATA UNDERSTANDING



before undersampling



after undersampling

TEXT PRE-PROCESSING

Text Preprocessing is a process to make dataset more "clean" so that more "ready" for model building.

What will be done in this process are:
1. Case folding, step to convert all characters into the same cases
2. Label encoding, step to convert target feature into numeric
3. Token masking, step to convert a specific "token" into a verbal form
4. Text cleaning, step to remove a specific "token" or "character" from text
5. Text normalization, step to normalize a text into the "normal" one
6. Stop words removal, step to stop words i.e. words/phrases that do not have meaning in our research from our text
7. Stemming, step to cut affixed words into a base form

TEXT PRE-PROCESSING

	Message_body	Label
0	If I die I want u to have all my stuffs.	Non-Spam
1	Hi.:)technical support.providing assistance to us customer through call and email:)	Non-Spam
2	But your brother transfered only <#> + <#> . Pa.	Non-Spam
3	Gd luck 4 ur exams :-)	Non-Spam
4	Wat time r ü going to xin's hostel?	Non-Spam

before text pre-processing

	Message_body	Label
0	die want u stuff	0
1	hitechn supportprovid assist us custom call email	0
2	brother transfer pa	0
3	good luck four ur exam	0
4	time ü go xin hostel	0

after text pre-procesing

FEATURE EXTRACTION

Process to select and/or combine variables into features. In text processing, feature extraction is done by converting text into a "number" using a "particular method"

Tokenization	Process to separate a text into smaller units called tokens
Count Vectorizer	Feature extraction technique that counting the total occurrences of term t in document d

FEATURE EXTRACTION

Message_body	Label
If I die I want u to have all my stuffs.	Non-Spam
Hi.:)technical support.providing assistance to us customer through call and email:)	Non-Spam
But your brother transfered only <#> + <#> . Pa.	Non-Spam
Gd luck 4 ur exams :-)	Non-Spam
Wat time r ü going to xin's hostel?	Non-Spam

before feature extraction

Message_body	aah	aare	aareedaare	aareel	aareeos	aberdeen	abl	abstract	abt	ac	...	yr	yrs	ys	yup	zebra	zed	zero	zf	zs	Label
get spiritu deep great	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
think ur smart win week week quiz text play phone number nowt winnersclub po box uz gbpweek	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
back good journey let know need receipt shall tell like pendent	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
great safe trip panic surrend	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
send resum	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 1793 columns

after feature extraction

MODELLING

Split dataset

The dataset is divided into training data and test data with a ratio of 80:20

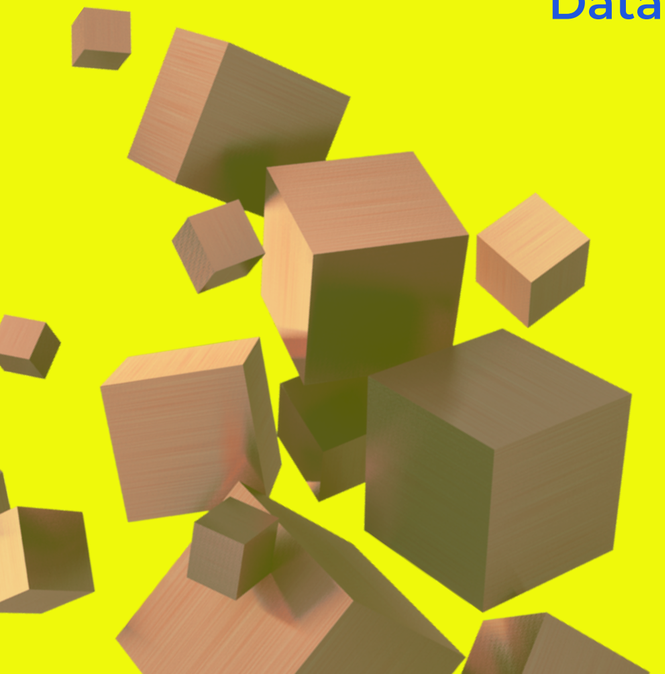
Build model

The models to be used are the Random Forest Classifier and Logistic Regression

MODELLING

Message_body	aah	aare	aareedaare	aareel	aareeos	aberdeen	abl	abstract	abt	ac	...	young	yr	yrs	ys	yup	zebra	zed	zero	zf	zs
wrong phone phone answer one assum peopl well	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
take away money worri	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
win winner mr foley ipod excit prize soon keep eye ur mobil visit websit numbercouk	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
want new nokia phone numberi colour phone deliveredtomorrow free minut mobil free text free camcord repli call phone number	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
guy plan come	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Data training with 418 records



Message_body	aah	aare	aareedaare	aareel	aareeos	aberdeen	abl	abstract	abt	ac	...	young	yr	yrs	ys	yup	zebra	zed	zero	zf	zs
shall bring us bottl wine keep us amus joke ill bring one anyway	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
free rington text first phone number poli text get phone number true tone help phone number phone number free tone x end txt stop	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
ok	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
gud gudk chikku tke care sleep well gud nyt	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
want gym	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Data testing with 105 records

EVALUATION

01

Accuracy score

02

Classification report

03

Confusion matrix

EVALUATION

Accuracy score: 96.23%

	precision	recall	f1-score	support
0	0.94	1.00	0.97	64
1	1.00	0.90	0.95	42
accuracy			0.96	106
macro avg	0.97	0.95	0.96	106
weighted avg	0.96	0.96	0.96	106

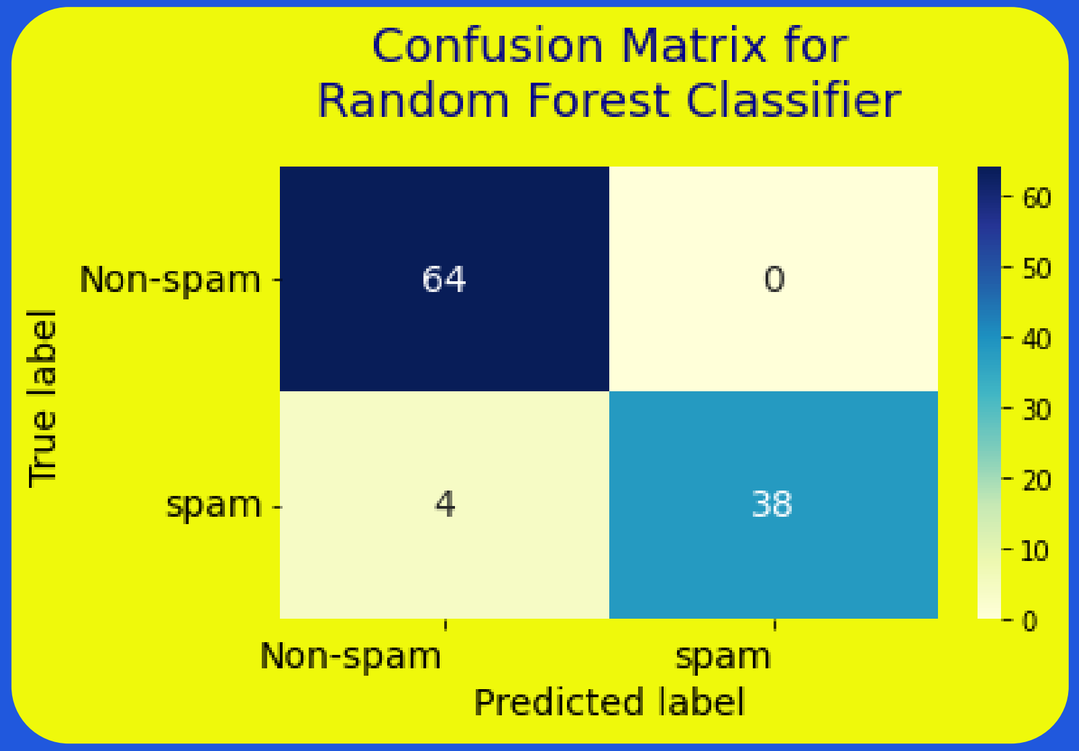
Accuracy score and classification report Random Forest

Accuracy score: 95.28%

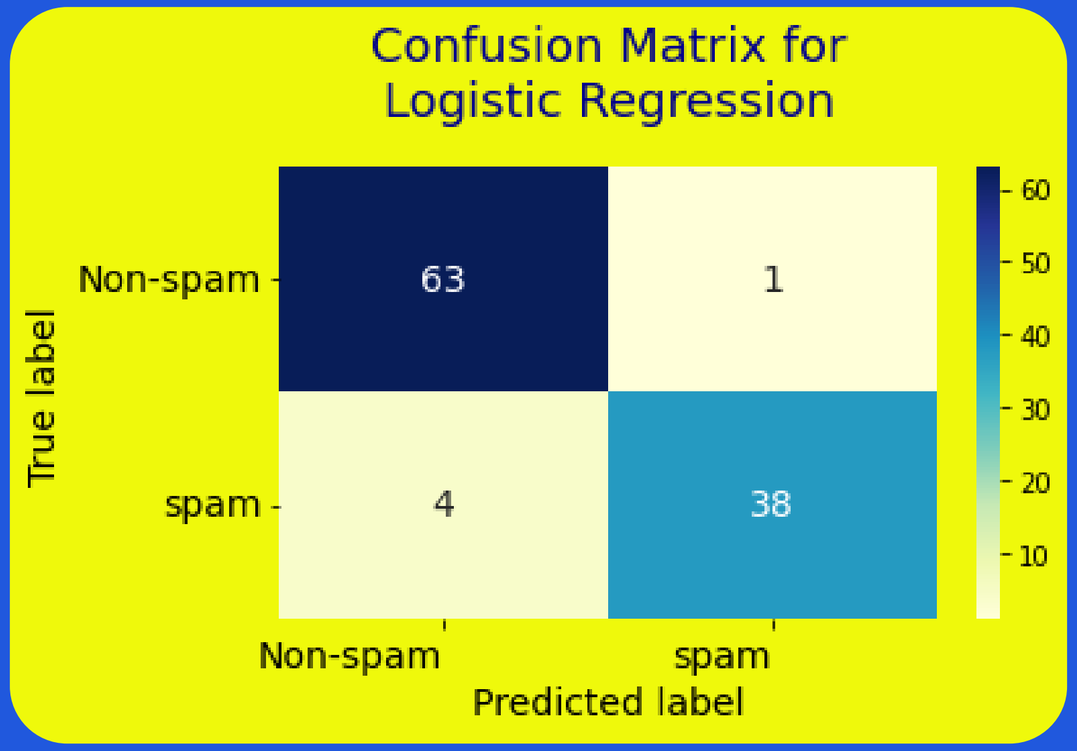
	precision	recall	f1-score	support
0	0.94	0.98	0.96	64
1	0.97	0.90	0.94	42
accuracy			0.95	106
macro avg	0.96	0.94	0.95	106
weighted avg	0.95	0.95	0.95	106

Accuracy score and classification report Logistic Regression

EVALUATION



Confusion matrix for
Random Forest



Confusion matrix for
Logistic Regression

SUMMARY

- The accuracy score for Random Forest Algorithm is `96.23% `
- The Accuracy score for Logistic Regression Algorithm is `95.28%`
- Random forest and logistic regression both have good accuracy for classifying labels. However, in this case, random forest has a slightly higher accuracy rate than logistic regression.

Kampus Merdeka INDONESIA JAYA

Bisa.ai PT. Bisa Artifisial Indonesia

THANK YOU

linkedin.com/in/andreanynthn

github.com/andreanynthn

and21yonathan@gmail.com