Kindle Book Review Classification

Senior Data Scientist Capstone Project

Dhananjaya 7th December 2021

Table of Contents

1

PROBLEM STATEMENT

User with most review, number of good rating, etc

2

DATA PREPROCESSING

Lowercase, punctuation removal, stopwords removal, etc

3

EDA

Exploratory Data Analysis

4

MODELLING

Building & Finding the best model

5

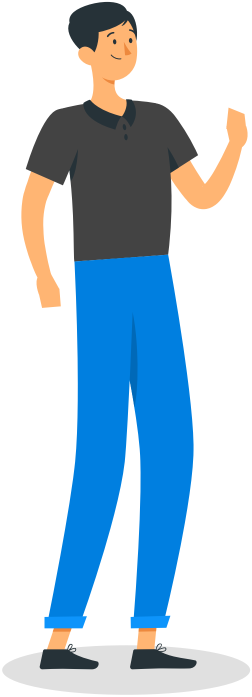
TESTING

Test the model using Amazon Book Review

6

SUMMARY

Summary of this project



1

Problem Statement

Dataset Review

Dataset yang dipakai adalah dataset Kindle Book Review. Yang mempunyai 12.000 record, dan 11 column. Dengan 4 column dengan type data int64, dan 7 column dengan type data object.

Unnamed: 0 (int64)	reviewTime (Object)
Unnamed: 0.1 (int64)	reviewerID (Object)
Asin (Object)	reviewerName (Object)
Helpful (Object)	Summary (Object)
Rating (int64)	unixReviewTime (int64)
reviewTex (Object)	

Problem Statement

a) Username yang memberikan paling banyak review ?
b) Tanggal yang mempunyai review paling banyak ?
c) Rating apa yang paling banyak ?
d) Bagaimana memprediksi rating dari review baru yang akan diberikan user ?
e) Model apa yang digunakan dan akurasi scorenya ?

Data Preprocessing

2

Data Preprocessing - Lowercase

Preprocessing 1 = mengubah text menjadi lowercase

Before :	Great short read. I didn't want to put it down so I read it all in one sitting.
After :	great short read. i didn't want to put it down so i read it all in one sitting.
Before :	I'll start by saying this is the first of four books so I wasn't expecting it to &#34;conclude&#34;. It centers
After :	i'll start by saying this is the first of four books so i wasn't expecting it to &#34;conclude&#34;. it centers

Data Preprocessing – Punctuation Removal

Preprocessing 2 = Punctuation Removal (. , ? ! : ; ‘ “)

Before :	great short read. i didn't want to put it down so i read it all in one sitting
After :	great short read i didn t want to put it down so i read it all in one sitting
Before :	i'll start by saying this is the first of four books so i wasn't expecting it to &#34;conclude&#34;. it centers
After :	i ll start by saying this is the first of four books so i wasn t expecting it to conclude it centers

Data Preprocessing – Stopwords Removal

Preprocessing 3 = Stopwords Removal

Before :	['great', 'short', 'read', 'i', 'didn', 't', 'want', 'to', 'put', 'it', 'down', 'so', 'i', 'read', 'it', 'all', 'in', 'one', 'sitting‘]
After :	['great', 'short', 'read, 'want', 'put', 'read', 'one', 'sitting‘]
Before :	['i', 'll', 'start', 'by', 'saying', 'this', 'is', 'the', 'first', 'of', 'four', 'books', 'so', 'i', 'wasn', 't', 'expecting', 'it', 'to', 'conclude', 'it', 'centers‘]
After :	['start', 'saying', 'first', 'four', 'books', 'expecting', 'conclude', 'centers‘]

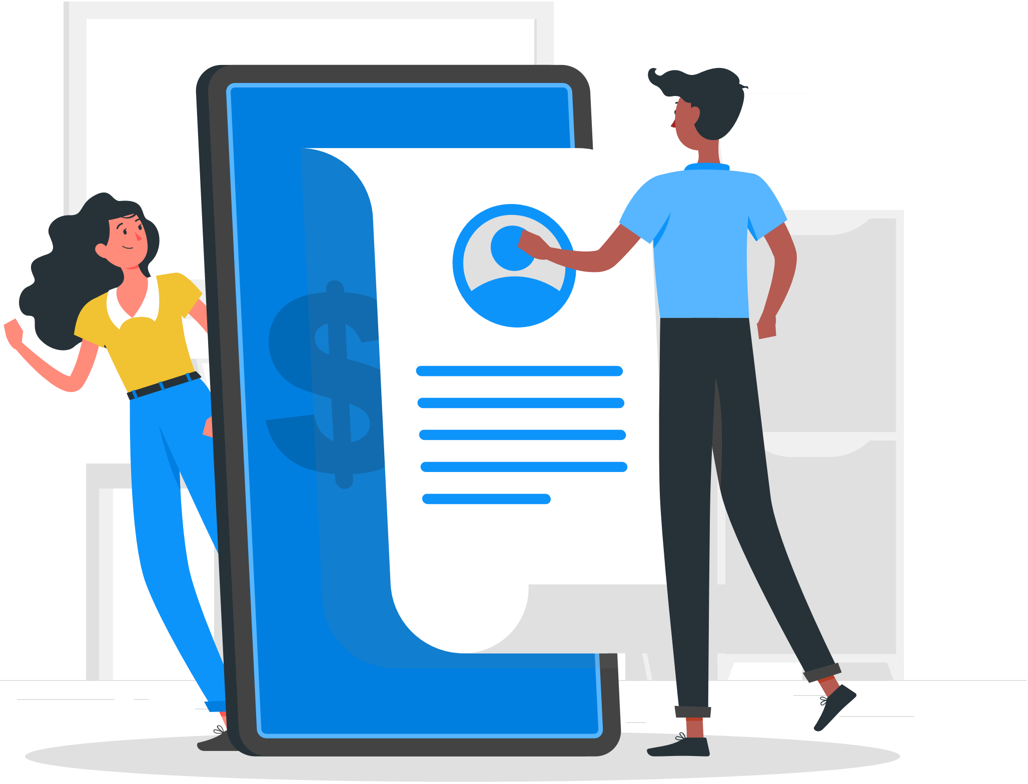
Data Preprocessing – Stemming

Preprocessing 4 = Stemming

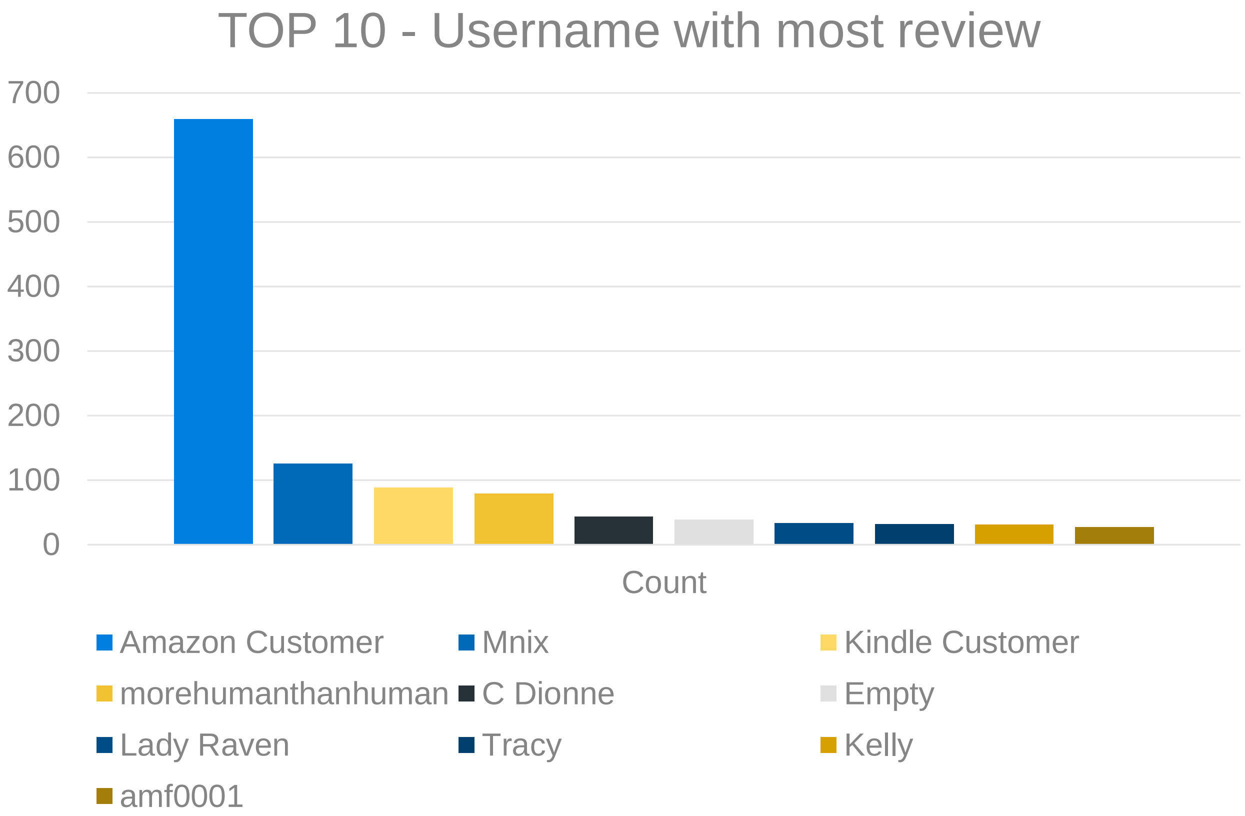
Before :	['great', 'short', 'read', 'want', 'put', 'read', 'one', 'sitting]
After :	['great', 'short', 'read, 'want', 'put', 'read', 'one', 'sit‘]
Before :	['start', 'saying', 'first', 'four', 'books', 'expecting', 'conclude', 'centers‘]
After :	['start', 'say', 'first', 'four', 'book', 'expect', 'conclud', 'center‘]

EDA

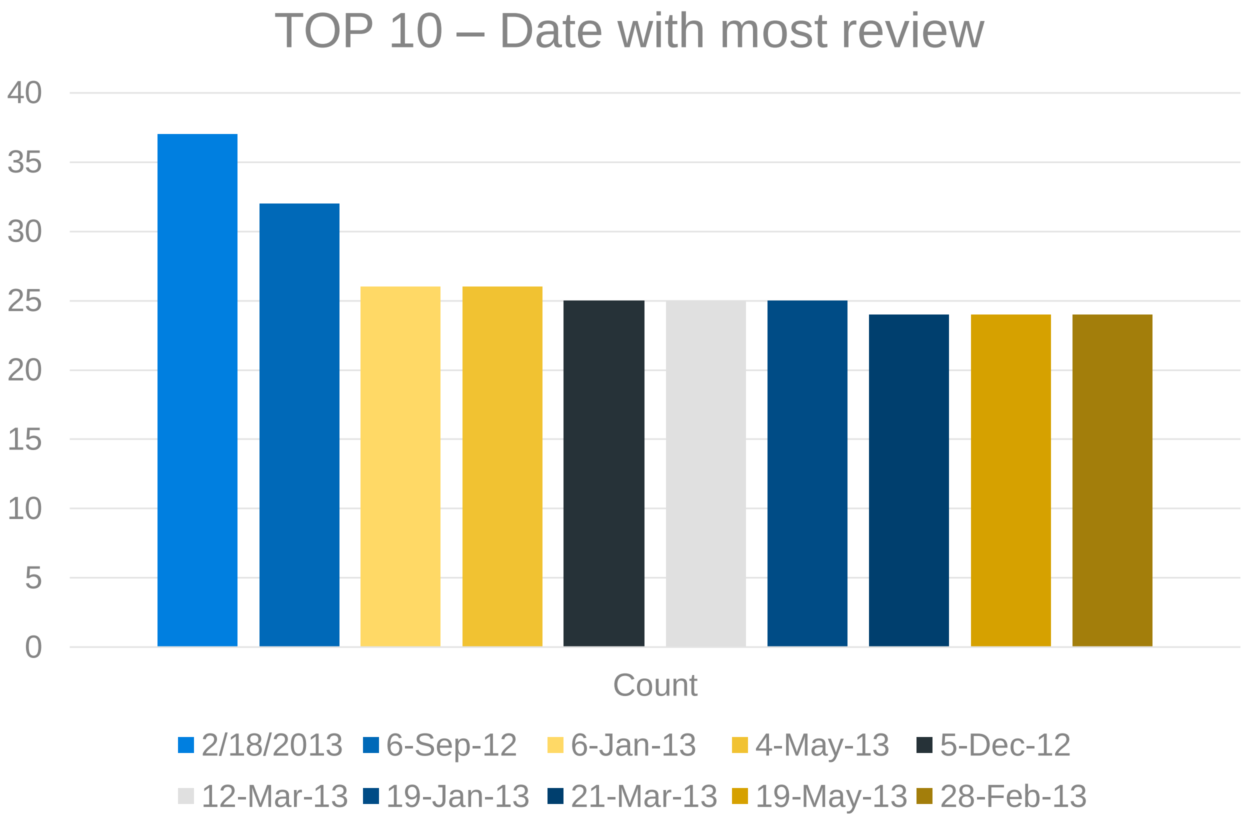
3



EDA - 1

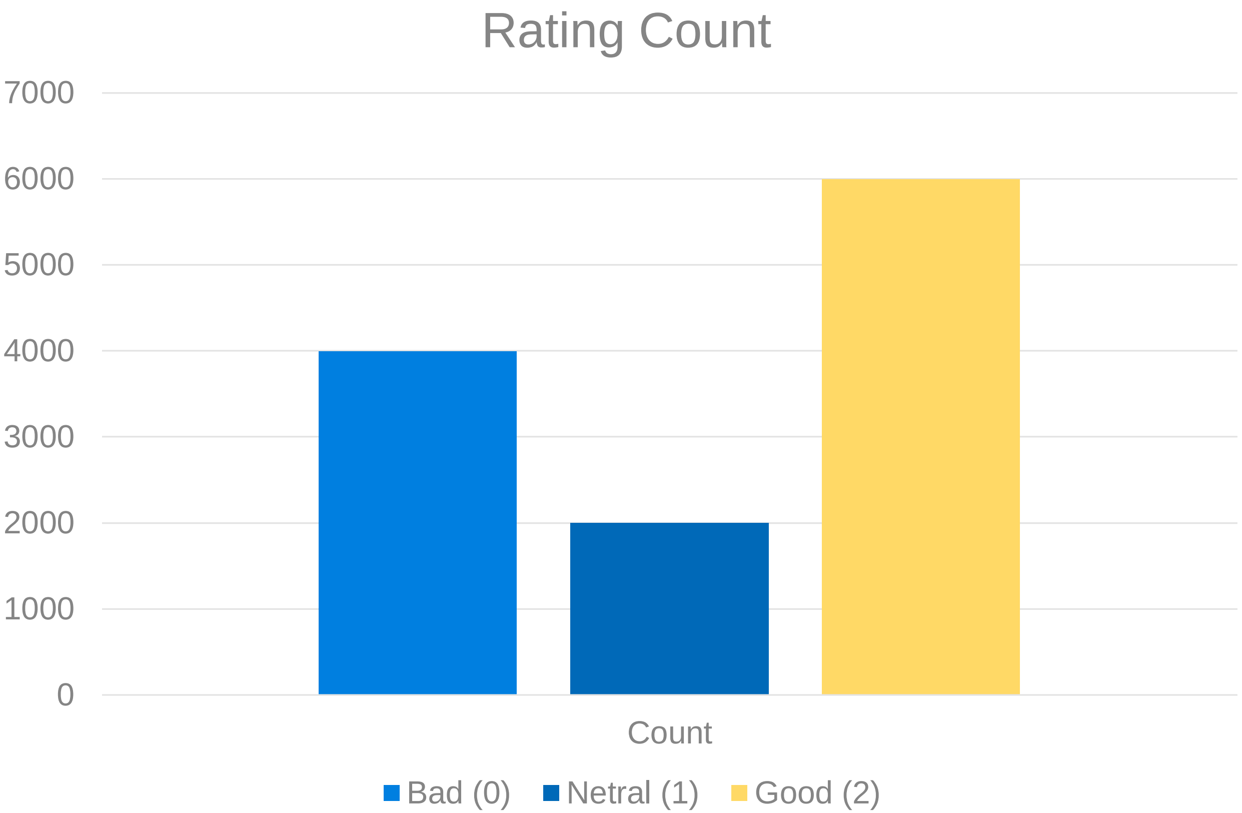

TOP 10 - Username with most review
700
600
500
400
300
200
100
0
Count
Amazon Customer
Mnix
Kindle Customer
morehumanthanhuman
C Dionne
Empty
Lady Raven
Tracy
Kelly
amf0001

EDA - 2



TOP 10 – Date with most review

EDA - 3


Rating Count
7000
6000
5000
4000
3000
2000
1000
0
Count
Bad (0)
Netral (1)
Good (2)

4

Modelling

Modelling

4

NGRAM

Unigram
BiGram
TriGram
UniGram & BiGram

4

ALGORITHM

Multinomial NaiveBayes
XGBoost
RandomForest
Support Vector Machine

16

MODEL COMPARISON

4 Ngram & 4 Algorithm

Model Comparison

	Unigram (1,1)	Bigram (2,2)	Trigram (3,3)	Unigram & Bigram (1,2)	AVG
Multinomial Naïve Bayes	65.8%	58.2%	49.9%	55.6%	**57.38%**
XGBoost	67.0%	57.5%	51.0%	67.8%	**60.83%**
Random Forest	70.3%	63.9%	52.6%	68.8%	**63.90%**
SVM	74.8%	69.8%	51.3%	75.8%	**67.93%**
AVG	**69.48%**	**62.35%**	**51.20%**	**67.00%**	

Model Comparison – Lowest Accuracy

	Unigram (1,1)	Bigram (2,2)	Trigram (3,3)	Unigram & Bigram (1,2)
Multinomial Naïve Bayes	65.8%	58.2%	49.9%	55.6%
XGBoost	67.0%	57.5%	51.0%	67.8%
Random Forest	70.3%	63.9%	52.6%	68.8%
SVM	74.8%	69.8%	51.3%	75.8%

Model Comparison – Highest Accuracy

	Unigram (1,1)	Bigram (2,2)	Trigram (3,3)	Unigram & Bigram (1,2)
Multinomial Naïve Bayes	65.8%	58.2%	49.9%	55.6%
XGBoost	67.0%	57.5%	51.0%	67.8%
Random Forest	70.3%	63.9%	52.6%	68.8%
SVM	74.8%	69.8%	51.3%	**75.8%**

5

Model Testing

Model Test With New User Review

Text	Predicted Rating
this book is very good. i want to read the book again. i recommend this book for you. best book ever	2 (correct)
this book is very bad. very disappointing story. i should never bought this book	0 (correct)
i can finish this book all night without sleeping	0 (wrong)
i want to read this book again and again and again	2 (correct)
the book content are very hard to understand	0 (correct)

Model Test With Amazon Book Review

Book Name : The Lincoln Highway : A Novel

Rating : 5 Star

Text	Real Rating	Predicted Rating
A long and winding road. It's 1954 and newly released from a work camp, Emmet wants nothing more than to pack up … etc	5 Star	2 (Correct)
I don't often read novels but, having devoured the author's two previous novels, I couldn't resist. And boy, am I glad I did. Such a simple story… etc	5 Star	2 (Correct)
I loved this Book!!!! The descriptions were fabulous and I could visualize everything just like I was watching a movie. The characters were wonderful and I especially loved Billy. …. etc	5 Star	2 (Correct)

https://www.amazon.com/product-reviews/B08WRH53MY/ref=acr_dp_hist_5?ie=UTF8&filterByStar=five_star&reviewerType=all_reviews#reviews-filter-bar

Model Test With Amazon Book Review 🧰

Book Name : The Lincoln Highway : A Novel

Rating : 3 Star

Text	Real Rating	Predicted Rating
Had this book been written by someone who I had never heard of, it would have received 4 stars. It is not a 5 star book no matter what, but it was written by the one … etc	3 Star	2 (Wrong)
Not as good as the Moscow book, but still an enjoyable read. The characters are a bit shallower and flatte… etc	3 Star	1 (Correct)
Towles writes so well that one cannot give him a truly bad review. And this book has moments of wonderful characterizations and scenes. However, I think he made a big mistake …. etc	3 Star	0 (Wrong)

https://www.amazon.com/product-reviews/B08WRH53MY/ref=cm_cr_unknown?ie=UTF8&filterByStar=three_star&reviewerType=all_reviews&pageNumber=1#reviews-filter-bar

Model Test With Amazon Book Review

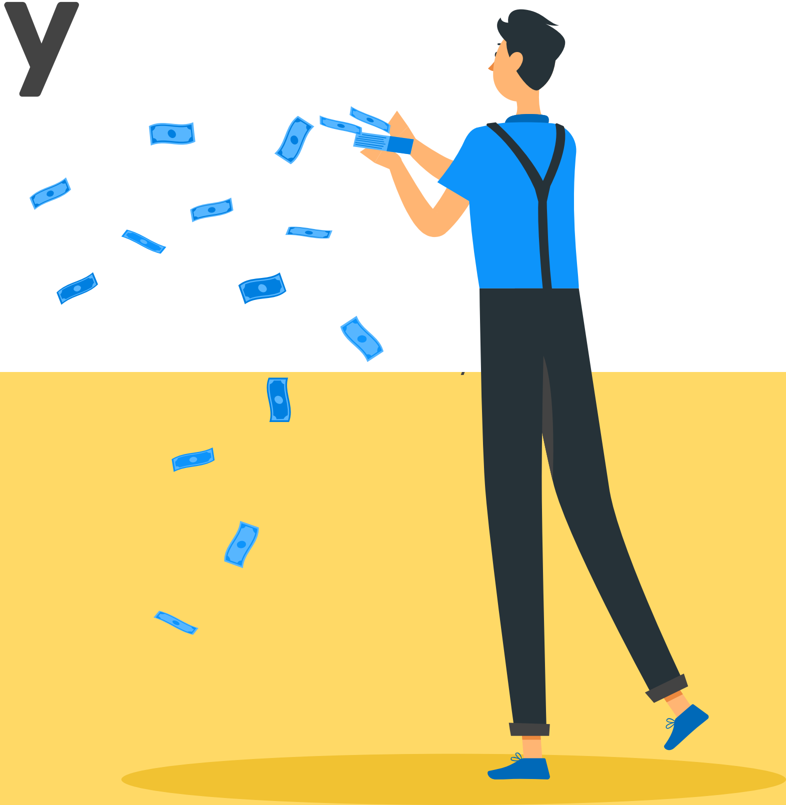
Book Name : The Lincoln Highway : A Novel

Rating : 1 Star

Text	Real Rating	Predicted Rating
The principles are teenage boys in the early 1950s--except for an eight-year-old little brother. The narrator ricochets from one of the characters to another, their thoughts (except, maybe, for Emmett's)... etc	1 Star	0 (Correct)
I thought his two first books were brilliant. This one is horrible. Uninteresting characters, no action, strange punctuation. Really hard to read.... etc	1 Star	0 (Correct)
I was so excited to learn that Towles had a third book and couldn't wait to read it. What a letdown after reading A Gentleman in Moscow and Rules of Civility. I found the book lacked interest and the characters .... etc	1 Star	0 (Correct)

https://www.amazon.com/product-reviews/B08WRH53MY/ref=cm_cr_unknown?ie=UTF8&filterByStar=one_star&reviewerType=all_reviews&pageNumber=1#reviews-filter-bar

6 Summary



Summary (1/2)

Results :

- Dari 4 jenis algoritma, dengan Avg accuracy paling rendah = Multinomial Naïve Bayes (Avg 57.38%).
- Dari 4 jenis algoritma, dengan Avg accuracy paling tinggi = SVM (Avg 67.93%)
- Dari 4 jenis ngram, dengan Avg accuracy paling rendah =Trigram (Avg 51.20%).
- Dari 4 jenis ngram, dengan Avg accuracy paling tinggi adalah Unigram (Avg 69.48%).
- Dari 16 jenis model, dengan accuracy paling rendah = Multinomial Naïve Bayes, ngram = Trigram, Accuracy = 49.9%
- Dari 16 jenis model, dengan accuracy paling tinggi =SVM, ngram = Unigram & Bigram, Accuracy = 75.8%

Summary (2/2)

Results :

- Dari hasil test menggunakan Amazon Book Review, diketahui Model bekerja dengan baik untuk memprediksi Rating 0 (bad)& 2 (good), tetapi tidak maksimal untuk memprediksi rating 1 (netral).
- Hyperparameter Tuning difokuskan ke metode SVM, tetapi Karena membutuhkan waktu training yang lama, sehingga hasilnya tidak maksimal. Dengan best parameter = {'C': 3, 'gamma': 1, 'kernel': 'sigmoid'}. Nilai accuracy = 75.1%. Nilai yang dihasilkan masih dibawah model standard. Sehingga model ini tidak menggunakan hyperparameter tuning.
- Penggunaan Hyperparameter Tuning dengan GridSearchCV harus dilakukan untuk meningkatkan model accuracy.

Thank You

slidesgo