Kotak Riset SC

ANALISIS SENTIMEN MENGGUNAKAN ALGORITMA RANDOM FOREST PADA BERITA INDONESIA

Rabbani Nur Kumoro
Audrey Shafira Fattima
William Hilmy Susatyo

Kotak Riset SC

Meet the Team

Bani

William

Audrey

Kotak Riset SC

Introduction

Tujuan

Mengkategorikan apakah suatu berita bersifat positif, negatif, ataupun netral berdasarkan data yang diberikan seperti judul, tokoh, jabatan, organisasi, dan kutipan.

Manfaat

Dapat melihat bagaimana sentimen media atau masyarakat terhadap suatu tokoh, jabatan, atau organisasi.

Kotak Riset SC

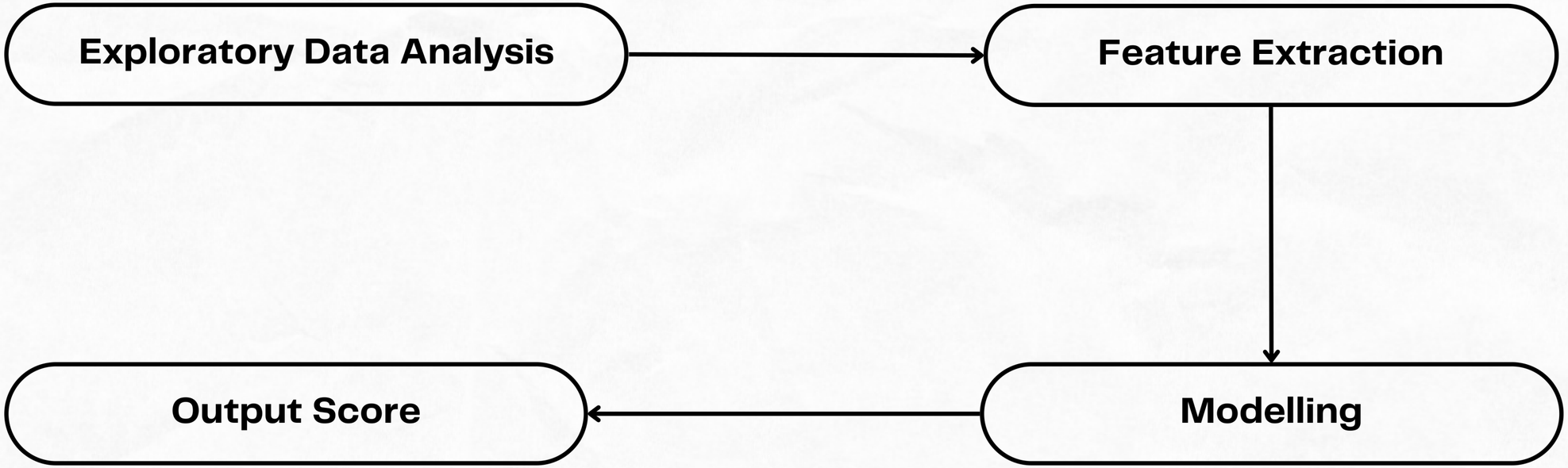
Data Awal

	id	sumber	kodekat	kodesubkat	kategori	subkategori	katakunci	tanggal	judul_berita	konten_berita	nama_tokoh	jabatan
0	00e3395ef29	Antara	J	J	Informasi dan Komunikasi	Informasi dan Komunikasi	Laporan keuangan	2021-02-05	ASN penyeleweng dana infak Masjid Raya divonis...	Padang (ANTARA) - Oknum Aparatur Sipil Negara ...	['Rinto']	['Hakim Ketua Yose']
1	019a47ed0bc	Detik	A	A2	Pertanian, Kehutanan, dan Perikanan	Kehutanan dan Penebangan Kayu	hasil hutan	2021-01-19	10 Alasan MK Kategorikan Ganja Hidroponik seba...	Jakarta - \n\nMahkamah Konstitusi (MK) memasukk...	['Dengler', 'Recommended Methods For', 'I']	[]
2	01eb3258ed4	Antara	D	D1	Pengadaan Listrik dan Gas	Ketenagalistrikan	Listrik PLN	2021-03-31	Angkasa Pura minta maaf atas mati listrik di ...	Padang (ANTARA) - PT Angkasa Pura II selaku pe...	['Ikhwan Wahyudi', 'Adi Lazuardi']	[]
3	02319ba7dbc	Okezone	J	J	Informasi dan Komunikasi	Informasi dan Komunikasi	Jumlah Penonton	2021-05-07	Liga 1 2021 Digelar dengan Kehadiran Penonton,...	JAKARTA ,Äì Direktur Utama PT LIB, Akhmad Hadi...	['Akhmad Hadian Lukita', 'Akhmad', 'Akhmad Had...	['Direktur Utama PT LIB', 'Presiden']
4	026dd5917f6	Detik	J	J	Informasi dan Komunikasi	Informasi dan Komunikasi	Laporan keuangan	2021-01-23	Bill Gates Kuasai Tanah Pertanian, Netizen Ket...	Jakarta - \n\nAda hal baru yang mungkin belum b...	['Offut', 'Gates', 'Michael Larson', 'Donald T...	[]

organisasi	lokasi	alias	kutipan	sentimen_kutipan	sentimen_berita
[]	['Padang', 'Sumbar', 'Provinsi Sumbar', 'Inspe...	['dibayar paling lama', '¬©']	['"Menyatakan terdakwa terbukti secara sah dan...	Negatif	Negatif
['Mahkamah Konstitusi (MK)', 'MK', 'KBBI: poho...	['Jakarta', 'Yogyakarta', 'Manual', '>>']	[]	[]	Netral	Negatif
['PT Angkasa Pura II', 'PLN', 'BIM', '***1***']	['Padang', 'Sumbar', 'Kota Padang', 'Kabupaten...	['¬©']	['"Kami menyampaikan permohonan maaf yang sebe...	Negatif	Negatif
['The Jakmania', 'Kementerian Pemuda dan Olahr...	['Indonesia', 'Jakarta']	[]	['"Kajian terus kami lakukan dan itu tidak gam...	Netral	Negatif
[]	['Jakarta', 'Amerika Serikat', 'Amerika', 'Afr...	[]	['"Tak boleh ada seseorang atau perusahaan bis...	Negatif	Negatif

Kotak Riset SC

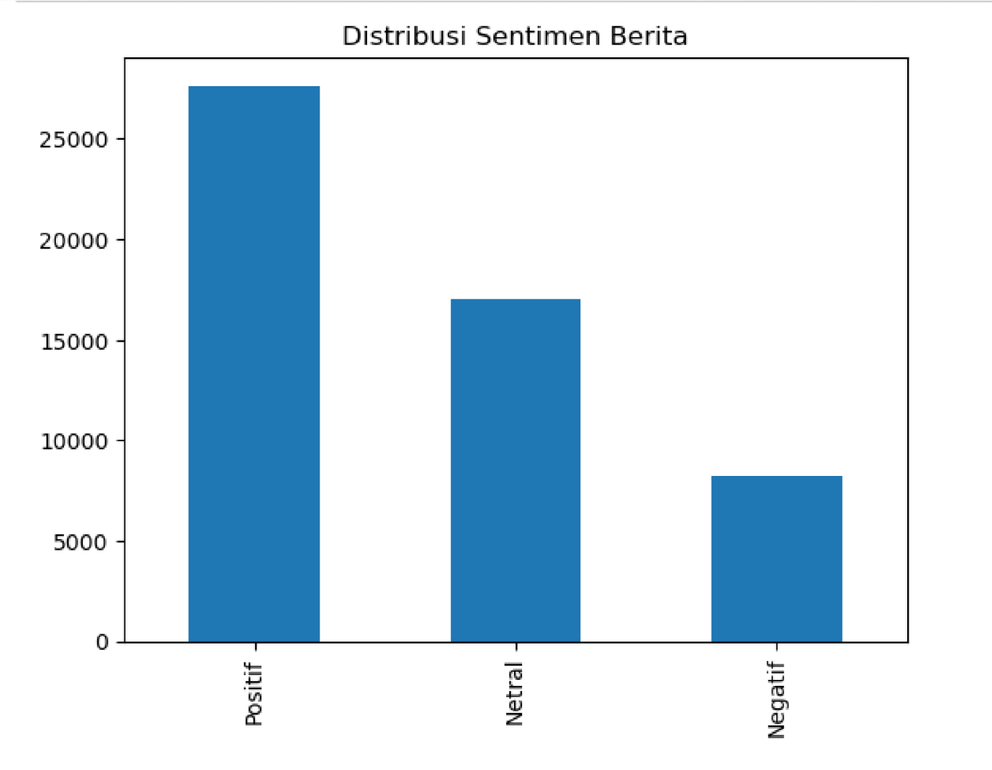
Alur Penelitian

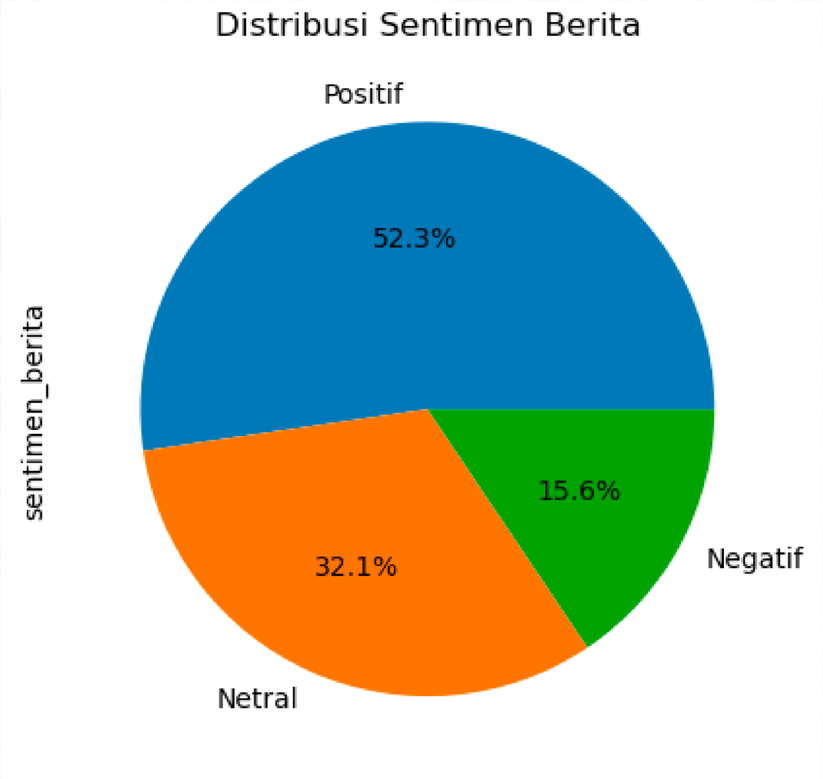


Kotak Riset SC

Alur Penelitian

EDA : Visualisasi

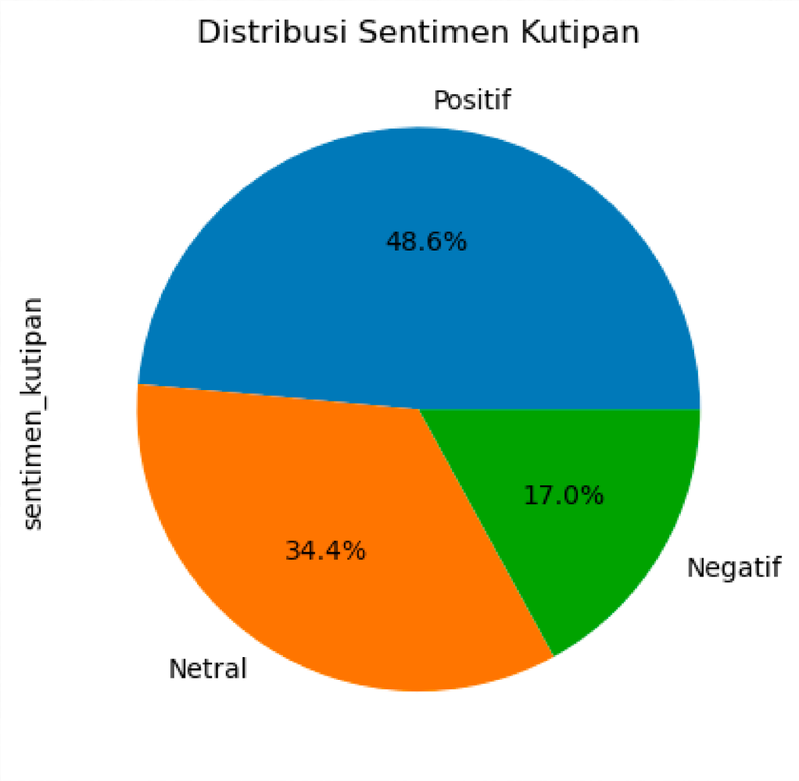


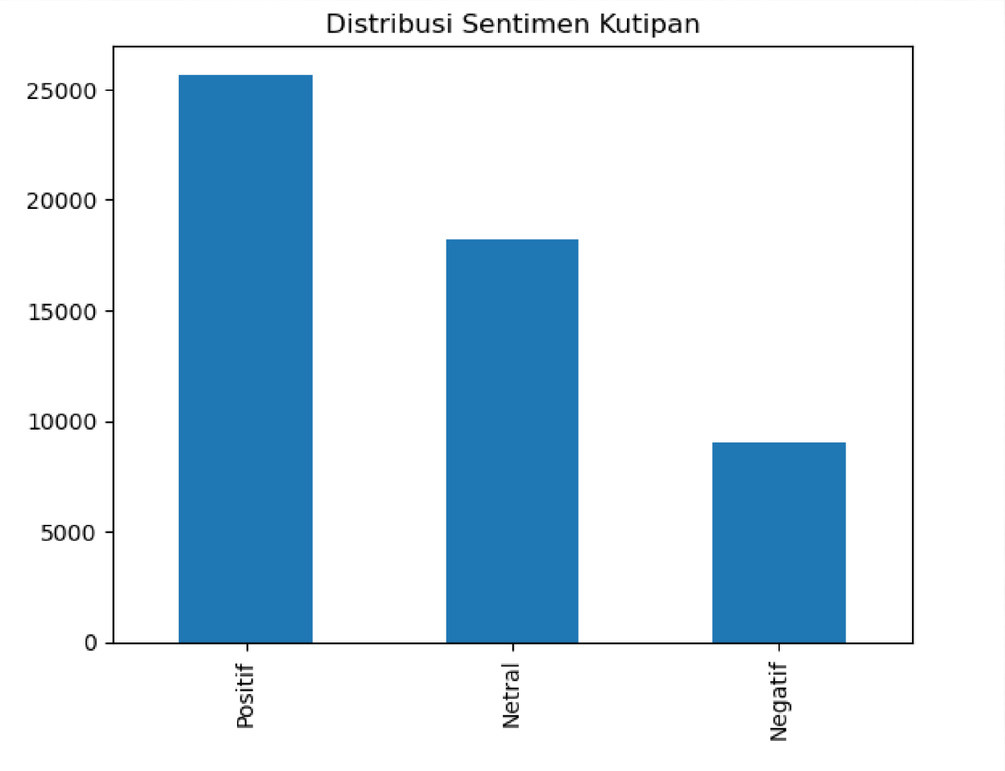


Kotak Riset SC

Alur Penelitian

EDA : Visualisasi

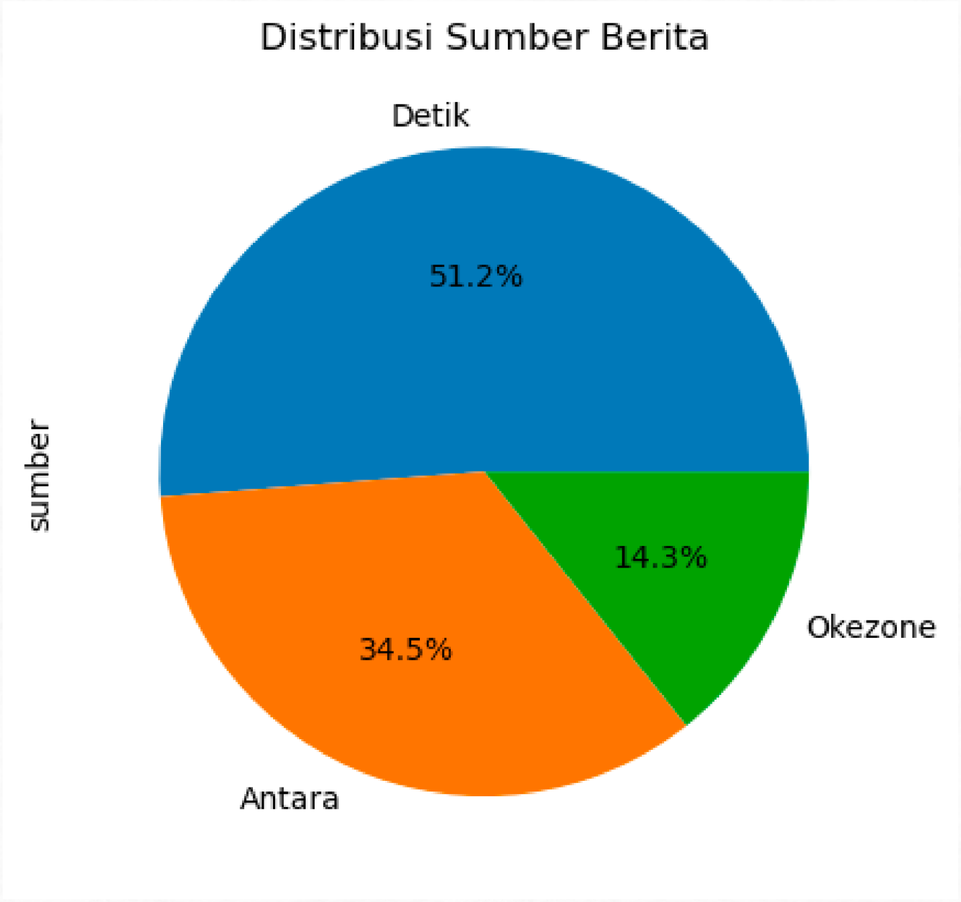




Kotak Riset SC

Alur Penelitian

EDA : Visualisasi



Kotak Riset SC

Alur Penelitian

EDA : Data Cleaning

```
def cleaningText(text):
    text = re.sub(r'@[A-Za-z0-9]+', '', text) # remove mentions
    text = re.sub(r'#[A-Za-z0-9]+', '', text) # remove hashtag
    text = re.sub(r'RT[\s]', '', text) # remove RT
    text = re.sub(r"http\S+", '', text) # remove link
    text = re.sub(r'[0-9]+', '', text) # remove numbers

    text = text.replace('\n', ' ') # replace new line into space
    text = text.translate(str.maketrans('', '', string.punctuation)) # remove all punctuations
    text = text.strip(' ') # remove characters space from both left and right text
    text = text.strip('[]')
    return text

def casefoldingText(text): # Converting all the characters in a text into lower case
    text = text.lower()
    return text

def tokenizingText(text): # Tokenizing or splitting a string, text into a list of tokens
    text = word_tokenize(text)
    return text

def stemmingText(text): # Reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of wo
    factory = StemmerFactory()
    stemmer = factory.create_stemmer()
    text = [stemmer.stem(word) for word in text]
    return text

def filteringText(text): # Remove stopwors in a text
    listStopwords = set(stopwords.words('indonesian'))
    filtered = []
    for txt in text:
        if txt not in listStopwords:
            filtered.append(txt)
    text = filtered
    return text

def toSentence(list_words): # Convert list of words into sentence
    sentence = ' '.join(word for word in list_words)
    return sentence
```

Kotak Riset SC

Clean Data

	id	sumber	kodekat	kodesubkat	kategori	subkategori	katakunci	tanggal	judul_berita	konten_berita	nama_tokoh	jabatan
0	00e3395ef29	Antara	J	J	Informasi dan Komunikasi	Informasi dan Komunikasi	Laporan keuangan	2021-02-05	asn penyeleweng dana infak masjid raya divonis...	padang oknum aparatur sipil negara asn pemprov...	rinto	hakim ketua yose
1	019a47ed0bc	Detik	A	A2	Pertanian, Kehutanan, dan Perikanan	Kehutanan dan Penebangan Kayu	hasil hutan	2021-01-19	alasan mk kategorikan ganja hidroponik pohon k...	jakarta mahkamah konstitusi mk memasukkan ganj...	dengler recommended methods for i	
2	01eb3258ed4	Antara	D	D1	Pengadaan Listrik dan Gas	Ketenagalistrikan	Listrik PLN	2021-03-31	angkasa pura maaf mati listrik bandara minangk...	padang pt angkasa pura ii pengelola bandara in...	ikhwan wahyudi adi lazuardi	
3	02319ba7dbc	Okezone	J	J	Informasi dan Komunikasi	Informasi dan Komunikasi	Jumlah Penonton	2021-05-07	liga digelar kehadiran penonton pt lib gampang	jakarta ,äì direktur utama pt lib akhmad hadia...	akhmad hadian lukita akhmad akhmad hadian joko...	direktur utama pt lib presiden
4	026dd5917f6	Detik	J	J	Informasi dan Komunikasi	Informasi dan Komunikasi	Laporan keuangan	2021-01-23	bill gates kuasai tanah pertanian netizen keta...	jakarta harta bill gates pendiri microsoft pem...	offut gates michael larson donald trump pengha...	

organisasi	lokasi	alias	kutipan	sentimen_kutipan	sentimen_berita
	padang sumbar provinsi sumbar inspektorat sumb...	dibayar ¬©	terdakwa terbukti sah bersalah tindak pidana k...	-1	-1
mahkamah konstitusi mk mk kbbi pohonpo¬∑hon n ...	jakarta yogyakarta manual			0	-1
pt angkasa pura ii pln bim	padang sumbar kota padang kabupaten padang	¬©	permohonan maaf sebesarbesarnya pengguna jasa ...	-1	-1
the jakmania kementerian pemuda olahraga kemen...	indonesia jakarta		kajian lakukan gampang mesti mencari referensi...	0	-1

Kotak Riset SC

Alur Penelitian

EDA : Visualisasi Clean Data

Kata-kata yang kerap disebutkan dalam Konten Berita

Kata-kata yang kerap disebutkan dalam Judul Berita

Kotak Riset SC

Alur Penelitian

Feature Extraction

Count Vectorizer

0	1	2	3	4	5	6	7	8	9	10	11
0	0	0	1	1	0	0	1	0	0	0	1
0	0	1	0	2	0	1	0	0	0	0	1
1	1	0	1	1	1	0	1	1	0	1	0

Sumber: tiksinc.online

TF-IDF Transformer

$$tfidf(w,d,D) = tf(w,d) * idf(w,D)$$

$$tf(w,d) = log(1 + f(w,d))$$

$$idf(w,D) = log(\frac{N}{f(w,D)})$$

Sumber: towardsdatascience

Kotak Riset SC

Alur Penelitian

Modeling

```
#Naive Bayes
naivebayes = MultinomialNB()
X_train_nb = np.nan_to_num(X_train)
y_train_nb = np.nan_to_num(y_train)
naivebayes.fit(X_train_nb, y_train_nb)
```

```
# Logistic Regression
logisticregression = LogisticRegression(solver='liblinear')
logisticregression.fit(X_train_nb, y_train_nb)
```

```
# Random Forest
from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_classification
randomforest = RandomForestClassifier()
randomforest.fit(X_train_nb, y_train_nb)
```

```
# Decision Tree
from sklearn.tree import DecisionTreeClassifier
decisiontree = DecisionTreeClassifier()
decisiontree.fit(X_train_nb, y_train_nb)
```

Kotak Riset SC

Result

F1 Score judul_berita vs F1 Score konten_berita

Model	F1 Score (judul_berita)	F1 Score (konten_berita)
Naive Bayes	0.6207	0.6308
Logistic Regression	0.6403	0.7352
Decision Tree	0.6639	0.7199
Random Forest	0.6932	**0.7372**

Kotak Riset SC

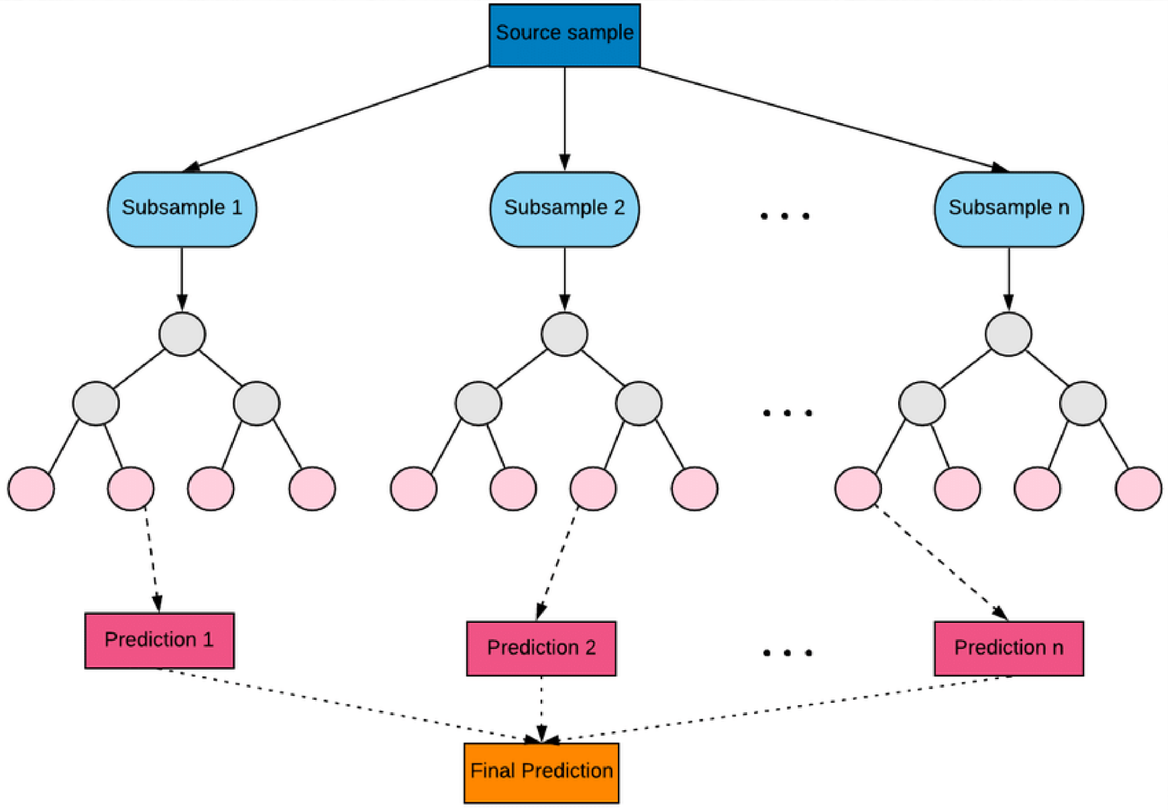
Metode

Random Forest

Why use Random Forest?

- Dapat digunakan untuk melakukan analisis regresi dan klasifikasi,
- Secara teoritis lebih akurat dibandingkan Decision Tree,
- Dapat melakukan analisis data dalam jumlah besar secara efisien.

How does Random Forest Works?



Sumber: researchgate.net

Kotak Riset SC

Final Result

Submission Kaggle

Submission and Description	Private Score	Public Score
submissionkotakrisetrf3.csv Complete · 3h ago	**0.75117**	**0.75619**
submissionkotakrisetrf2.csv Complete · 3h ago	**0.53147**	**0.53016**
submissionkotakrisetrf.csv Complete · 4h ago	**0.7075**	**0.70822**

Kotak Riset SC

Hartstikke
Bedankt!

Syukron!

Merci!

Gracias!

Thank You!

Terima Kasih!

Matur Suwun!