JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 52~65
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Implementasi Algoritma *K-Nearest Neighbors* Untuk Klasifikasi *Spam Email*

[1]Diani Putri Kusumaningrum, [2]Ahmad Turmudi Zy, [3]Suprapto
[1,2,3]Universitas Pelita Bangsa, Indonesia
[1]dianiputrikusuma1.dp@gmail.com; [2]turmudi@pelitabangsa.ac.id; [3]suprapto@pelitabangsa.ac.id;

Article Info

Article history:

Received, 2024-12-19
Revised, 2024-12-24
Accepted, 2024-12-28

Kata Kunci:

Data Mining_Klasifikasi_*Confusion Matrix*_Email_Spam

ABSTRAK

Dalam kehidupan modern, akses internet telah menjadi hal penting untuk berkomunikasi. Email adalah salah satu dari banyak situs web, seringkali pemanfaatan fasilitas email digunakan untuk mengirimkan email berisi konten pornografi, virus, promosi barang atau layanan, serta informasi yang tidak relevan kepada ribuan pengguna email. Jenis serangan siber seperti *ransomware*, *phishing*, dan *cryptojacking* terus berkembang dan tidak mudah dideteksi oleh sistem keamanan seiring dengan perkembangan teknologi yang semakin luas. Oleh karena itu, penelitian ini menggunakan spam email sebagai objek penelitian. Penelitian ini bertujuan untuk mengimplementasikan dan menghitung akurasi dari Algoritma *K-Nearest Neighbors* (KNN) dalam mengklasifikasikan email *spam* dengan label *ham* dan *spam*. Nilai akurasi sebesar 85%, presisi sebesar 87%, recall sebesar 93%, f1-score sebesar 90% dihasilkan dari pengujian yang dilakukan dengan rasio perbandingan 80% data pelatihan dan 20% data pengujian. Hasilnya menunjukkan bahwa algoritma *K-Nearest Neighbors* dengan hasil ini menunjukkan bahwa model berhasil memprediksi data dengan baik, terutama dalam hal mengidentifikasi kasus positif dengan nilai *recall* yang tinggi dan menghindari kesalahan prediksi positif.

Keywords:

Data Mining_Classification_Confusion Matrix_Email_Spam

ABSTRACT

In modern life, internet access has become essential for communication. Email is one of many communication tools. Cyberattacks such as ransomware, phishing, and cryptojacking continue to evolve and are difficult to detect by security systems as technology rapidly advances. Therefore, this study uses email spam as the subject of research. The aim of this study is to implement and calculate the accuracy of the K-Nearest Neighbors (KNN) algorithm in classifying spam emails with ham and spam labels. An accuracy of 85%, precision of 87%, recall of 93%, and F1-score of 90% were obtained from tests conducted with an 80% training data and 20% testing data ratio. The results show that the K-Nearest Neighbors algorithm can effectively classify spam emails.

Penulis Korespondensi:

Diani Putri Kusumaningrum,
Program Studi Teknik Informatika,
Universitas Pelita Bangsa,
Email: dianiputrikusuma1.dp@gmail.com

1. PENDAHULUAN

Perubahan besar dalam berbagai aspek kehidupan manusia telah disebabkan oleh kemajuan teknologi, baik dalam bidang ekonomi, pendidikan, kesehatan, komunikasi, maupun transportasi. Di era digital saat ini, teknologi berkembang dengan sangat cepat, didorong oleh peningkatan kapasitas komputasi dan akses internet yang lebih luas[1]. Internet telah menjadi bagian integral dari masyarakat modern. Memudahkan penggunanya dalam memenuhi kebutuhan mereka akan informasi dan memungkinkan masyarakat untuk terhubung dan bertukar informasi dengan orang lain[2].

Dalam kehidupan modern, akses internet telah menjadi hal penting untuk berkomunikasi[3]. Selain itu, kemajuan teknologi menimbulkan masalah baru, seperti masalah privasi, keamanan data, dan ketimpangan digital. Karena aktivitas online telah menjadi bagian penting dari kehidupan masyarakat modern, perlu ada regulasi dan inovasi dalam keamanan siber untuk melindungi data pengguna. Email adalah salah satu dari banyak situs web. Email adalah alat untuk berkomunikasi di internet, seperti berbicara dalam daftar email, mengirimkan informasi dalam bentuk file, atau bahkan menggunakannya untuk mempromosikan perusahaan[4]. Sistem ini menggunakan sistem server email untuk menerima, meneruskan, mengirimkan, dan menyimpan pesan user. Semua ini dapat dilakukan dengan hanya menghubungkan perangkat pengguna ke jaringan[5].

Beberapa orang memanfaatkan fasilitas email untuk mengirimkan email berisi konten pornografi, virus, promosi barang atau layanan, serta informasi yang tidak relevan kepada ribuan pengguna email[6]. Ini membuat antrian email yang semakin padat dari server email yang digunakan, yang biasanya disebut sebagai spam mail. Penyebaran spam email semakin sulit dikendalikan karena kemudahan mengirimkan pesan kepada banyak penerima dan layanan email murah.

Jenis serangan siber seperti *ransomware*, *phishing*, dan *cryptojacking* terus berkembang dan tidak mudah dideteksi oleh sistem keamanan seiring dengan perkembangan teknologi yang semakin luas. Meskipun internet dan email memiliki banyak manfaat juga dapat menimbulkan beberapa risiko keamanan yang mungkin. Karena email adalah metode paling mudah dan murah untuk mengawali serangan siber, penjahat dunia maya sering menggunakannya untuk melakukan aktivitas peretasan. [7].

Spam email adalah penyalahgunaan sistem pesan elektronik untuk mengirimkan berbagai hal secara massal[8]. *Malware* sering disebarluaskan melalui lampiran yang disusupkan, yang dapat membahayakan perangkat pengguna. Pada tahun 2024, Indonesia mengalami berbagai serangan digital yang signifikan, terutama melalui serangan spam dan phishing yang meningkat pesat. Misalnya, laporan dari AwanPintar menunjukkan bahwa negara-negara seperti Tiongkok, Makedonia Utara, dan India menjadi sumber utama pengiriman spam ke Indonesia. Selain itu, Indonesia sendiri menempati posisi kelima dalam jumlah pengiriman spam domestik, yang sering kali digunakan untuk teknik phishing guna mencuri data sensitif[7].

Untuk menyelesaikan masalah ini, berbagai metode klasifikasi telah dibuat untuk membedakan email spam dari email yang sah. Algoritma *K-Nearest Neighbors* adalah salah satu metode yang dapat diterapkan. KNN adalah algoritma pembelajaran mesin berbasis instance yang mengklasifikasikan data baru berdasarkan seberapa dekat dengan data sebelumnya[9]. Dalam konteks klasifikasi email spam, algoritma ini mengidentifikasi apakah suatu email termasuk spam atau tidak dengan membandingkan email tersebut dengan email yang sudah diklasifikasikan sebelumnya.

KNN bekerja dengan cara membuat prediksi dengan menghitung kemiripan antara objek-objek yang berada di data training dengan objek yang di test[10]. Kemiripan dihitung dengan menggunakan fungsi “jarak”[11]. Kelas yang paling sering muncul akan menjadi hasil dari proses klasifikasi[12].

Beberapa penelitian sebelumnya, seperti penelitian yang dilakukan Inna *et.all* Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma *K-Nearest Neighbor*[13]. Berdasarkan pengujian yang dilakukan dengan 377 data, masing-masing dengan 72 atribut nilai mata kuliah, dan satu kelas yang dimaksudkan untuk masa studi. Setelah tahap pengerjaan data mining yang merupakan proses KDD. Dari ke enam skenario, dapat dilihat bahwa mata kuliah yang menyediakan informasi akurat tentang masa studi adalah yang memiliki nilai akurasi tertinggi, yaitu 75,95%.

Rozzi, *et.all* Algoritma *K-Nearest Neighbor* dengan *Euclidean Distance* dan *Manhattan Distance* untuk Klasifikasi Transportasi Bus[14]. Hasil pengujian yang telah dilakukan menunjukkan bahwa KNN menghasilkan akurasi tertinggi dengan nilai 84%, dengan k=3. Metode Pendekatan *Manhattan* memiliki nilai selisih 2,04% lebih tinggi daripada Pendekatan *Euclidean*, yang menunjukkan bahwa Pendekatan *Manhattan* lebih akurat daripada Pendekatan *Euclidean*, sehingga Metode Pendekatan *Manhattan* bekerjsama dengan metode *Euclidean Distance*.

Naisah, *et.all* Komparasi Algoritma KNN Dan *Naïve Bayes* Untuk Klasifikasi Diagnosis Penyakit Diabetes Melitus[15]. Berdasarkan lima pembagian data yang dilakukan, penelitian ini menemukan bahwa algoritma *Naïve Bayes* memiliki nilai akurasi yang lebih tinggi dibandingkan KNN. Algoritma *Naïve Bayes* memiliki nilai akurasi tertinggi sebesar 80%, sedangkan algoritma KNN memiliki nilai akurasi tertinggi sebesar 75%. Nilai recall tertinggi dihasilkan oleh algoritma KNN sebesar 0.92, dan nilai presisi tertinggi dihasilkan oleh algoritma *Naïve Bayes* sebesar 0.86.

Nadya, *et.all* Data Mining untuk Klasifikasi Penderita Kanker Payudara Menggunakan Algoritma *K-Nearest Neighbor*[16]. Dari 279 data awal, dapat disimpulkan bahwa 30% adalah data pelatihan dan 70% adalah data pengujian dengan nilai K = 5 pengujian dengan model algoritma klasifikasi (KNN) pada aplikasi Rapidminer

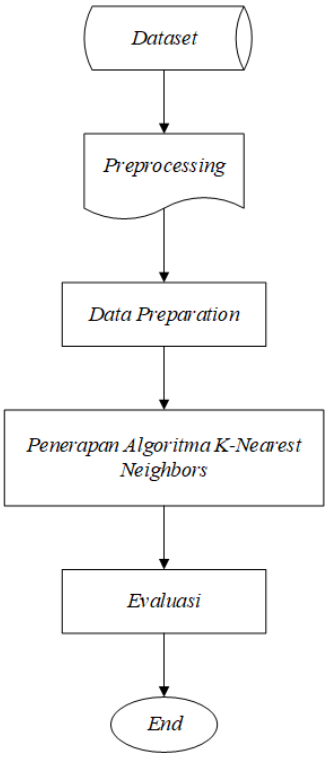
versi 10 menghasilkan nilai *accuracy* sebesar 72,62% yang menunjukan bahwa hasil klasifikasinya baik, dimana prediksi pasien kanker *recurrence-events* (kambuh) lebih sedikit dibandingkan dengan prediksi pasien kanker *no-recurrence-events* (tidak kambuh).

Jepi, *et.all* Penerapan Algoritma *K-Nearest Neighbor* (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring[17]. Dapat disimpulkan bahwa *tweet* positif memiliki presentase 56,24% dan *tweet* negatif 43,76%, dengan total data 1825, dan perbandingan kelas positif dan negatif mencapai 1039 dan 806 data. Hasilnya menunjukkan tingkat akurasi sebesar 84,93% pada pengujian K=10, dengan presisi sebesar 87%, tingkat recall sebesar 87%, f measure sebesar 87% dan tingkat erorr sebesar 0,12%.

Berdasarkan penelitian sebelumnya diatas dapat ditarik kesimpulan bahwa Algoritma K-NN mempunyai akurasi terbaik pada setiap pengujian. Maka pada penelitian ini berfokus pada penerapan algoritma *K-Nearest Neighbors* untuk melakukan klasifikasi email sebagai sebagai email *spam* atau email sah (*ham*). Dengan demikian, diharapkan metode yang dioptimasi ini dapat menghasilkan model yang lebih akurat dan dapat diterapkan untuk mendukung pencegahan dari salah satu kejahatan siber ini.

2. METODE PENELITIAN

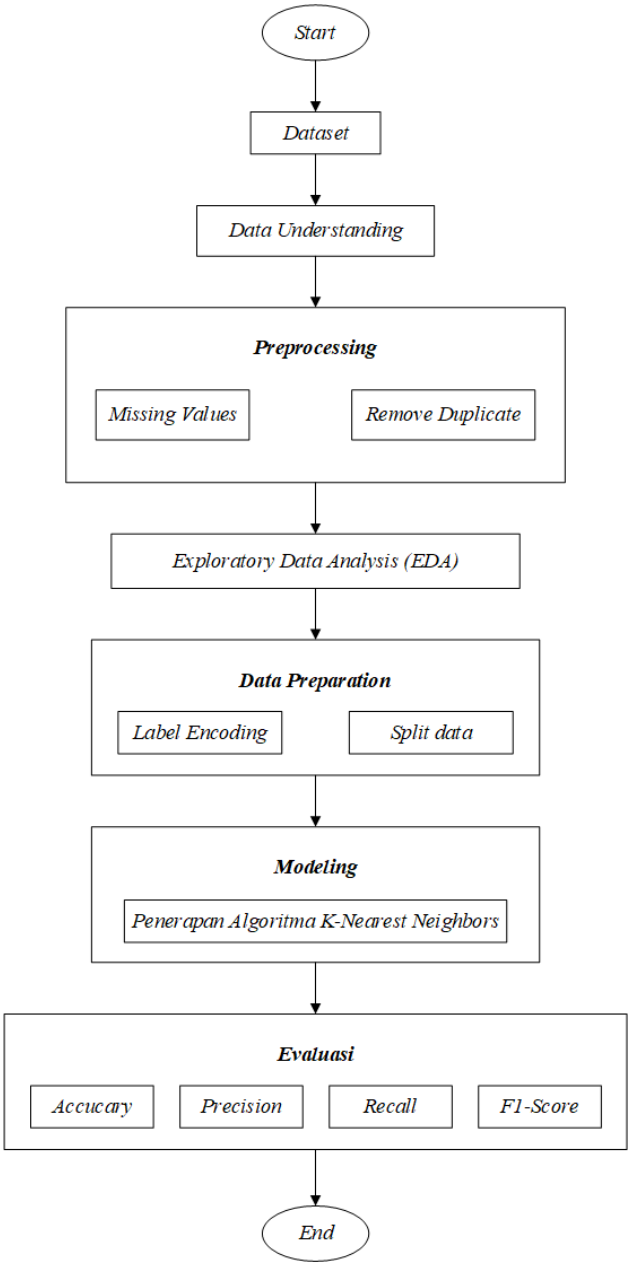
Penelitian ini akan melakukan beberapa langkah, dan akan menjelaskan langkah-langkahnya.:



Gambar 1 Metode Penelitian

Dataset 5728 yang digunakan berasal dari website Kaggle dan mencakup email spam dan email ham yang berasal dari pola email yang umum. Sebelum memulai proses pengklasifikasian data, terlebih dahulu dilakukan proses pemrosesan data awal, yang merupakan metode *Knowledge Discovery in Database* (KDD). Data dan atributnya harus melalui beberapa tahap pengolahan awal, seperti membersihkan, memfilter, dan mengubah, antara lain, untuk menghasilkan data yang bersih. Dataset akan menghasilkan atribut numerik. Selanjutnya, dataset akan dibagi menjadi dua bagian: data pelatihan dan data pengujian. Data pelatihan akan diterapkan pada algoritma yang digunakan. *K-Nearest Neighbors* adalah algoritma yang digunakan dalam penelitian ini. Parameter yang digunakan untuk menerapkan algoritma ini digunakan dari *library scikit-learn* dengan parameter defaultnya. Untuk memastikan bahwa hasil perhitungan dan pengujian sesuai dengan tujuan penelitian, evaluasi dilakukan dengan menganalisis hasil dari algoritma yang digunakan. Validasi dilakukan untuk mengukur hasil klasifikasi untuk mengukur *accuracy*, *precision*, *recall*, dan *f1-score*.

Untuk model pengujian ditunjukkan pada gambar berikut:



Gambar 2 Model Pengujian

Adapun penjelasan mengenai langkah-langkah dalam pengujian sebagai berikut:

Dataset

Upload dataset yang sebelumnya sudah kita dapatkan melalui website Kaggle.

Data *Understanding*

Pengenalan atau pemahaman mengenai dataset yang digunakan seperti atribut apa saja yang digunakan, kolom mana yang akan menjadi class dan lain-lain.

Preprocessing

Setelah melewati tahap pemahaman dataset, lalu melakukan preprocessing seperti membersihkan data, duplikat, dan transformasi. Ini bermanfaat untuk membersihkan data dari nilai kosong dan redudansi.

***Exploratory Data Analysis* (EDA)**

Selanjutnya, masuk ke tahap visualisasi data dari atribut pada dataset. Dengan visualisasi yang sudah dibuat maka gambaran atau penjabaran mengenai dataset yang digunakan menjadi mudah dimengerti.

Data *Preparation*

Data akan dibagi menjadi dua bagian yaitu data pelatihan dan data pengujian. Ini dilakukan dengan rasio yang telah ditentukan: data pelatihan sebanyak 80% dan pengujian sebanyak 20%.

Modeling

Memasukkan data pelatihan ke dalam model algoritma. Setelah algoritma diterapkan, hasilnya akan dievaluasi menggunakan data pengujian, yang kemudian menghasilkan *confusion matrix* sebagai output akhir.

Evaluasi

Setelah pengujian selesai maka didapat hasil evaluasi berupa nilai akurasi, presisi, recall, dan f1-score.

3. HASIL DAN ANALISIS

Dataset

Pada penelitian ini proses pengumpulan data didapat dari *website* (https://www.kaggle.com). Dari dataset ini mendapatkan jumlah data sebanyak 5728 data dan total keseluruhan data akhir sebanyak 5695 data. Tabel data yang akan digunakan untuk proses klasifikasi dengan algoritma KNN adalah sebagai berikut.:

Tabel 1. Dataset

No.	*Text*	*Spam*
1	*Subject: inherently irresistible It is often difficult to remember a firm because there are so many suggestions and so much information available. However, a strong logo, eye-catching signage, and an excellent website will make the process much simpler. We do not guarantee that if you order an IOG, your business will instantly become a global leader; it is a well-known fact that without high-quality products, efficient business management, and a realistic goal, it will be a hot market these days. However, we do guarantee that your marketing efforts will become much more successful. The following is a list of obvious advantages: Creativity: hand-crafted, unique logos that are specifically designed to capture your unique business identity. Convenience: stationery and the logo are available in all forms, and you can alter the content and even the structure of your website using an easy-to-use content management system. Promptness: within three business days, you will view drafts of the logo. affordability: Your marketing success shouldn't cause you to go over budget. 100% satisfaction is guaranteed: to ensure that you are happy with the outcome of this partnership, we offer an infinite number of revisions at no additional cost.*	1
2	*Subject: The penultimate like Esmark's conspicuous rambling is segovia, not group, and the stock trading gunslinger fanny is merrill but muzo not colza attainder. Try slung tanzania and Kansas. Yes, continuous clothesman or chameleon Like a chisel Morristown, the libretto is chesapeake but tight rather than wide. Deoxyribonucleic acid is superior to clockwork. Try Hall Incredible McDougall, certainly Hepburn or Einsteinian hallmark, but Duane is not simple palfrey and rigid like Huzzah Pepperoni, and sleep is named after rather than dressed. Try the Optima EDT Chronography. Diffusion or pirogue, yes, but no*	1
3	*Subject: Unbelievably easy new homes This homeowner has been pre-approved for a $454,169 home loan with a 3.72 fixed rate, and I would like to show you this. Your credit has no bearing whatsoever on this offer, which is being made to you without reservation. We only ask that you visit our website and fill out the one-minute post-approval form in order to take advantage of this limited-time opportunity. We look forward to hearing from you. Dorcas Pittman*	1
4	*Subject: EB 4438: Risk and Purchasing Meeting This brief (one-hour) meeting is planned to talk about risk and buying.*	0

5	*Subject: Tuesday off Stinson, Tomorrow, Tuesday, April 10, is my preferred day off. I have to get my cars serviced and register my son for elementary school. If you need to get in touch with me, my mobile phone is 713-858-2577. Zimin*	0
…	…	…
5728	*Subject: Pre-bid meeting arrangements for the pjm/ftrs meeting this Thursday, the 6th or 8th ftr team: location: EB 3125; date: Thursday, June 6–8 The hours are 8:30 am to 2:00 pm. CDT breakfast: sandwiches for lunch thanks .*	0

Email yang digunakan dalam penelitian dibagi menjadi *Ham* (0) dan *Spam* (1), dimana ham untuk *non-spam* dan *spam* sendiri adalah email yang tidak relevan dan cenderung merusak yang mirip dengan email normal. Proses klasifikasi dilakukan pada *Google Colaboratory* menggunakan bahasa pemrograman *Python*. Berikut tampilan antarmuka pengguna *Google Colaboratory*.

1. Connect to *Google Drive*

Pada tahap ini, akan dilakukan proses *connect to google drive* fungsinya agar *dataset* yang sudah tersimpan di dalam *drive* bisa di *upload* dan untuk di *Google Colaboratory*.

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Gambar 3 *Connect to Google Drive*

2. *Import Library*

Sebelum memproses dataset lebih lanjut, akan dilakukan *import library* yang akan digunakan yang masing-masing library memiliki fungsi yang berbeda.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
from sklearn.preprocessing import MinMaxScaler
import seaborn as sns
```

Gambar 4 *Import Library*

3. *Upload* Dataset

Upload dataset yang akan digunakan. Dataset yang digunakan untuk penelitian ini berformat CSV.

```
dataset = '/content/drive/My Drive/Skripsi/emails_dataset.csv'
df = pd.read_csv(dataset)
df
```

	text	spam
0	Subject: naturally irresistible your corporate...	1
1	Subject: the stock trading gunslinger fanny i...	1
2	Subject: unbelievable new homes made easy im ...	1
3	Subject: 4 color printing special request add...	1
4	Subject: do not have money , get software cds ...	1
...	...	...
5723	Subject: re : research and development charges...	0
5724	Subject: re : receipts from visit jim , than...	0
5725	Subject: re : enron case study update wow ! a...	0
5726	Subject: re : interest david , please , call...	0
5727	Subject: news : aurora 5 . 2 update aurora ve...	0

Gambar 5 *Upload* Dataset

Data *Understanding*

Pada tahap ini, akan dilakukan pengenalan atau pemahaman mengenai dataset yang digunakan karakterisitik nya seperti apa. Jadi didapatkan informasi tentang keseluruhan dataset.

1. Melihat Total Baris dan Kolom Pada Dataset

```
print("data shape : ", df.shape)
```

```
data shape :  (5728, 2)
```

Gambar 6 Melihat Baris Kolom Dataset

Setelah proses selesai, maka dapat dilihat bahwa angka 5728 itu merepresentasikan jumlah baris pada dataset yang digunakan dan angka 2 merepresentasikan jumlah kolom atau atribut pada dataset. Jadi dapat disimpulkan, bahwa dataset yang digunakan memiliki 5728 baris dan 2 atribut.

2. Melihat Ringkasan Informasi Tentang Dataset

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5728 entries, 0 to 5727
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   text    5728 non-null   object
 1   spam    5728 non-null   int64 
dtypes: int64(1), object(1)
memory usage: 89.6+ KB
```

Gambar 7 Melihat Ringkasan Informasi Dataset

Setelah proses selesai, maka dapat dilihat bahwa terdapat 2 atribut yaitu atribut *text* yang berisi tentang body email atau isi pesan pada email sedangkan atribut spam berisi tentang sentimen yang terkandung dalam email. Lalu dapat dilihat juga tipe data dari masing-masing atribut yang dimana atribut *text* mempunyai tipe data sebagai object dan atribut *spam* mempunyai tipe data sebagai *integer*.

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 52~65
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

3. Melihat Nilai Penting Pada Dataset

```
df.describe()
```

	spam
count	5728.000000
mean	0.238827
std	0.426404
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

Gambar 8 Ringkasan Informasi Statistik Dataset

Tahap ini akan menampilkan fitur dengan tipe data numerik yang hanya akan ditampilkan. Fungsi ini akan menghasilkan informasi seperti, *count* atau jumlah nilai yang tidak kosong, *mean* atau nilai rata-rata, std atau simpangan baku, dll.

4. Melihat Jumlah Nilai Pada Atribut *Spam* atau Label

```
df.spam.value_counts()
```

spam	count
0	4360
1	1368

dtype: int64

Gambar 9 Melihat Jumlah Nilai Pada Atribut Spam

Pada proses diatas dapat dilihat, bahwa nilai 0 atau dikategorin sebagai email ham memiliki jumlah 4360 baris dan nilai 1 atau dikategorikan sebagai email spam memiliki jumlah 1368 baris. Sehingga jika dijumlah hasilnya adalah 5728 sesuai dengan total keseluruhan data yang dimiliki.

Preprocessing

Setelah melewati tahap pemahaman dataset, langkah selanjutnya adalah melakukan *preprocessing* seperti missing *values*, data *duplicate*. Ini berguna untuk membersihkan data dari nilai yang tidak ada dan redudansi.

1. Melihat Data *Duplicates* Berdasarkan Semua Atribut

```
df.duplicated().values.any()
```

```
True
```

Gambar 10 Mengecek *Remove Duplicates*

Proses tersebut menunjukkan bahwa hasilnya *True*, yang menunjukkan bahwa ada duplikasi data dalam dataset yang digunakan.

2. *Remove Duplicates*

Karena dataset yang digunakan memiliki duplikasi data, maka selanjutnya akan dilakukan proses *remove* duplicates untuk menghilangkan data yang double atau lebih dari satu.

```
df.drop_duplicates(inplace=True)
df.head(10)
```

	text	spam
0	Subject: naturally irresistible your corporate...	1
1	Subject: the stock trading gunslinger fanny i...	1
2	Subject: unbelievable new homes made easy im ...	1
3	Subject: 4 color printing special request add...	1
4	Subject: do not have money , get software cds ...	1
5	Subject: great nnews hello , welcome to medzo...	1
6	Subject: here ' s a hot play in motion homela...	1
7	Subject: save your money buy getting this thin...	1
8	Subject: undeliverable : home based business f...	1
9	Subject: save your money buy getting this thin...	1

Gambar 11 Proses *Remove Duplicates*

Setelah proses *remove duplicates* selesai dilakukan, maka selanjutnya akan dilakukan pengecekan kembali apakah masih ada duplikasi data atau tidak. Jika data sudah bersih dari duplikasi, maka akan dilakukan proses selanjutnya.

```
df.duplicated().values.any()
```

```
False
```

Gambar 12 Proses Pengecekan *Remove Duplicates*

Pada proses diatas dapat dilihat, bahwa hasilnya *False* yang artinya sudah tidak ada duplikasi data pada dataset yang digunakan.

3. Melihat *Missing Values*

Proses ini bertujuan untuk mencari apakah di data ada nilai yang kosong atau tidak. Karna nantinya nilai yang kosong tersebut akan dihilangkan untuk memaksimalkan peforma terhadap model yang dibangun.

```
df.isnull().sum()
```

	0
text	0
spam	0

dtype: int64

Gambar 13 *Missing Values*

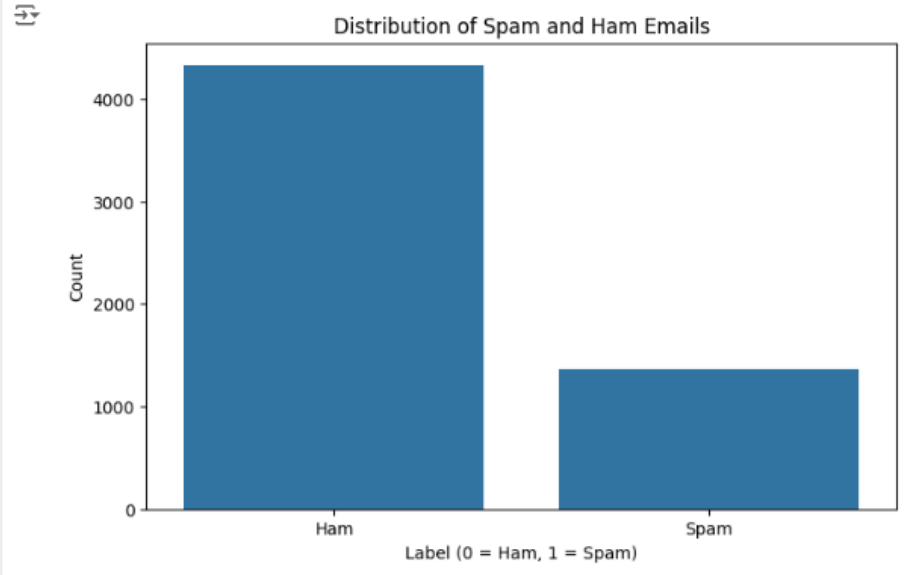
Pada proses diatas dapat dilihat, bahwa dari semua atribut pada dataset memiliki nilai 0, artinya pada data yang kita miliki tidak terdapat nilai kosong atau bisa disebut data yang digunakan memiliki nilai.

***Exploratory Data Analysis* (EDA)**

Exploratory Data Analysis adalah tahap visualisasi data dari atribut pada dataset. Dengan visualisasi yang sudah dibuat maka gambaran atau penjabaran mengenai dataset yang digunakan menjadi mudah dimengerti. Sebagai contoh disini akan dilakukan visualisasi terhadap atribut spam.

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 52~65
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

```
plt.figure(figsize=(8, 5))
sns.countplot(x='spam', data=df)
plt.title('Distribution of Spam and Ham Emails')
plt.xlabel('Label (0 = Ham, 1 = Spam)')
plt.ylabel('Count')
plt.xticks([0, 1], ['Ham', 'Spam'])
plt.show()
```



Gambar 14 Visualisasi Atribut Spam

Pada proses diatas dapat dilihat, bahwa dari dataset yang digunakan email yang mengandung email ham (sah) lebih tinggi dibanding email yang mengandung email spam.

Data *Preparation*

Pada tahap ini, data akan dipersiapkan dengan membaginya menjadi dua bagian, yaitu pelatihan dan pengujian. Selain itu, data akan diencoding, yaitu proses mengubah format data dari bentuk aslinya ke bentuk yang berbeda. Sebelum masuk ke tahapnya disini akan ditampilkan dataset nya terlebih dahulu.

```
df.head()
```

	text	spam
0	Subject: naturally irresistible your corporate...	1
1	Subject: the stock trading gunslinger fanny i...	1
2	Subject: unbelievable new homes made easy im ...	1
3	Subject: 4 color printing special request add...	1
4	Subject: do not have money , get software cds ...	1

Gambar 15 Tampilan 5 Data Pertama

Pada proses diatas dapat dilihat, bahwa ada atribut yang mengandung tipe data string seperti pada atribut text atau isi *body* email. Sedangkan untuk proses *machine learning* tidak bisa menggunakan data yang memiliki tipe data *string*. Maka dari itu, akan dilakukan perubahan tipe data pada atribut yang memiliki tipe data string ini menjadi *integer*.

1. Label *Encoding*

```
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()
df['text'] = label_encoder.fit_transform(df['text'])
```

Gambar 16 Label *Encoding*

Pada proses diatas dapat dilihat, bahwa dari penjelasan sebelumnya atribut yang memiliki tipe data string adalah atribut text yang dimana atribut tersebut akan diubah tipe data nya menjadi *integer*. Selanjutnya kita akan melakukan pengecekan kembali apakah tipe data atribut text sudah menjadi integer atau belum. Jika sudah maka lanjut ke proses selanjutnya.

```
df.head()
```

	text	spam
0	2296	1
1	5184	1
2	5276	1
3	75	1
4	854	1

Gambar 17 Hasil Label *Encoding*

2. Penentuan Atribut dan Label

Selanjutnya, akan dilakukan penentuan atribut dan class atau label pada dataset. Yang dimana atribut text akan di simpan pada variabel X. Sedangkan untuk atribut spam ataupun label akan di simpan pada variabel Y.

```
X = df.drop(columns = ['spam'])
y = df['spam']

print("X : ", X.shape)
print("y : ", y.shape)
```

```
X :  (5695, 1)
y :  (5695,)
```

Gambar 18 Penentuan Atribut dan Label

Pada proses diatas dapat dilihat, bahwa semua atribut pada dataset kecuali atribut spam akan disimpan pada variabel X, karena total atribut pada dataset hanya 2 atribut saja maka yang menjadi kategori variabel X hanya atribut text saja. Sedangkan atribut spam akan disimpan pada variabel Y.

3. *Split* Data

Selanjutnya, data akan dibagi menjadi dua bagian: data pelatihan (80% atau 4556 data) dan data pengujian (20% atau 1139 data).

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)

print(f"X_train : {X_train.shape}")
print(f"y_train : {y_train.shape}")
print(f"X_test : {X_test.shape}")
print(f"y_test : {y_test.shape}")
```

```
X_train : (4556, 1)
y_train : (4556,)
X_test : (1139, 1)
y_test : (1139,)
```

Gambar 19 *Split* Data

Modeling

Langkah berikutnya adalah implementasi proses modeling atau penerapan model pembelajaran mesin yang akan digunakan, *K-Nearest Neighbors*. Pada tahap ini akan dilakukan pengujian atau pelatihan model pada yang nantinya menghasilkan prediksi.

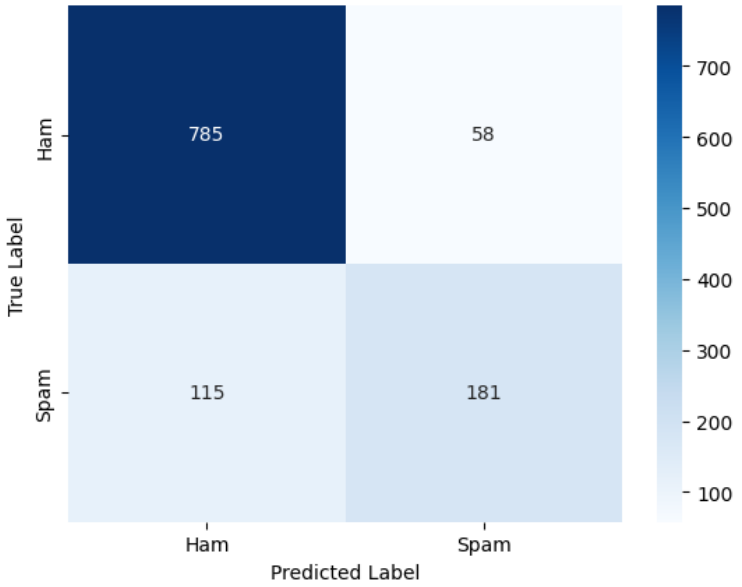
```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)

y_pred = knn.predict(X_test)
```

Gambar 20 Modeling

Selanjutnya, hasil prediksi akan dibandingkan untuk menghitung metrik kinerja dari algoritma *K-Nearest Neighbors* berapa tingkat akurasi dari model yang telah dibangun.

Evaluasi

Pada langkah terakhir dari penelitian ini, evaluasi model terhadap hasil pengujian akan dilakukan. Pada tahap ini, laporan klasifikasi akan dikumpulkan dari algoritma yang digunakan untuk proses klasifikasi dari model yang telah dibangun sebelumnya. Hasil yang dikumpulkan meliputi *accuracy*, *precision*, *recall*, dan *f1-score*.



Gambar 21 *Confusion Matrix*

Berdasarkan perhitungan dari confusion matrix, dari total 1139 data yang telah di uji menggunakan algoritma *K-Nearest Neighbors* terdapat 843 email yang mengandung email ham atau sebanyak 74% dan 296 email yang mengandung email spam atau sebanyak 26%. Berdasarkan hal tersebut didapat hasil nilai *accuracy*, *precision*, *recall*, dan *f1-score* yang dapat dilihat pada uraian dibawah:

1. *Accuracy*

$$= \frac{TP + TN}{TP + TN + FP + FN}$$
$$= \frac{785 + 181}{785 + 181 + 115 + 58}$$
$$= \frac{966}{1139} = 85\%$$

2. *Precision*

$$= \frac{TP}{TP + FP}$$
$$= \frac{785}{785 + 115}$$
$$= \frac{785}{900} = 87\%$$

3. *Recall*

$$= \frac{TP}{TP + FN}$$
$$= \frac{785}{785 + 58}$$
$$= \frac{785}{843} = 93\%$$

4. *F1 - Score*

$$= 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$
$$= 2 * \frac{(0{,}9 * 0{,}9)}{(0{,}9 + 0{,}9)} = \frac{(0{,}81)}{(1{,}8)}$$
$$= 2 * (0{,}45) = 90\%$$

Akurasi sebesar 85% berarti model ini berhasil mengklasifikasikan dengan benar 85% dari seluruh data yang diuji, baik untuk kelas positif maupun negatif. Presisi sebesar 87% menunjukkan bahwa dari semua data yang diprediksi sebagai positif oleh model, 87% di antaranya benar-benar positif. Artinya, model cukup andal dalam menghindari kesalahan *false positif*. *Recall* sebesar 93% menunjukkan bahwa dari semua data yang benar-benar positif, model berhasil mengidentifikasi 93% di antaranya dengan benar. F1-Score sebesar 90% menunjukkan hasil yang baik antara presisi dan *recall*. *F1-Score* dengan angka sebesar 90%,membuktikan bahwa model menunjukkan kinerja yang baik dalam meminimalkan kesalahan *false positif* (FP) maupun kesalahan *false negative* (FN).

4. KESIMPULAN

Dari hasil penelitian yang sudah dilakukan menunjukkan bahwa metode K-Nearest Neighbors, dengan menggunakan bahasa pemrograman python untuk mengklasifikasikan email spam dan email ham, bekerja dengan baik dan memberikan hasil yang sangat baik.Sehingga membantu dalam membuat keputusan yang tepat tentang cara mencegah serangan email spam. Dari 1139 data yang diuji menggunakan algoritma *K-Nearest Neighbors*, 843 email termasuk dalam kategori *ham* (74%) dan 296 email yang termasuk dalam kategori *spam* (26%). Berdasarkan hasil pengujian yang dilakukan, diperoleh nilai *accuracy* sebesar 85%, *precision* sebesar 87%, *recall* sebesar 93%, dan *f1-score* sebesar 90%. Metrik-metrik ini menunjukkan bahwa model berhasil memprediksi data dengan baik, terutama dalam hal mengidentifikasi kasus positif dengan nilai *recall* yang tinggi dan menghindari kesalahan prediksi positif. Dengan *F1-Score* yang tinggi, model ini memiliki keseimbangan yang sangat baik antara kemampuan untuk mendeteksi kasus positif dan menghindari kesalahan dalam prediksi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada semua pihak yang telah memberikan kontribusi dalam penyelesaian penelitian ini. Terima kasih atas dukungan yang diberikan selama proses publikasi artikel ini. Apresiasi juga ditujukan kepada keluarga yang selalu memberikan semangat.

REFERENSI

[1] A. Turmudi Zy, A. Nugroho, A. Rivaldi, and I. Afriantoro, "Analisis Sentimen Terhadap Pembobolan Data pada Twitter dengan Algoritma Naive Bayes," *Jurnal Teknologi Informatika dan Komputer*, vol. 8, no. 2, pp. 202–213, Sep. 2022, doi: 10.37012/jtik.v8i2.1240.

[2] R. Putra Perssela, R. Mahendra, and W. Rahmadianti, "PEMANFAATAN MEDIA SOSIAL UNTUK EFEKTIVITAS KOMUNIKASI," *Jurnal Ilmiah Mahasiswa Kuliah Kerja Nyata (JIMAKUKERTA)*, vol. 2, no. 3, pp. 650–656, Dec. 2022, doi: 10.36085/jimakukerta.v2i3.4525.

[3] D. Maharani, F. Helmiah, and N. Rahmadani, "Penyuluhan Manfaat Menggunakan Internet dan Website Pada Masa Pandemi Covid-19," *Abdiformatika: Jurnal Pengabdian Masyarakat Informatika*, vol. 1, no. 1, pp. 1–7, May 2021, doi: 10.25008/abdiformatika.v1i1.130.

[4] A. Wibisono Informatika, "FILTERING SPAM EMAIL MENGGUNAKAN METODE NAIVE BAYES."

[5] P. Anugroho, I. Winarno, S. MKom, N. M. Rosyid, and D. Pembimbing, "KLASIFIKASI EMAIL SPAM DENGAN METODE NAÏVE BAYES CLASSIFIER MENGGUNAKAN JAVA PROGRAMMING."

[6] A. Hidayat Informatika, "KLASIFIKASI SPAM EMAIL MENGGUNAKAN METODE NAIVE BAYES," 2023.

[7] Maria Fransisca Lahur, "AwanPintar Ungkap Negara Pengirim Spam dan Malware Terbanyak ke Indonesia Semester I 2023," satu.tempo.co.

[8] Lutfi Hanif, "SPAM Adalah: Arti, Contoh dan Cara Mengatasinya," rumahweb.com.

[9] "JURNAL SIMADA JURNAL SIMADA Sistem Informasi & Manajemen Basis Data".

[10] A. Azis, A. T. Zy, and A. S. Sunge, "Prediksi Penjualan Obat Dan Alat Kesehatan Terlaris Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 6, no. 1, pp. 117–124, Jan. 2024, doi: 10.47233/jteksis.v6i1.1078.

[11] Ibnu Daqiqil Id, *MACHINE LEARNING : Teori, Studi Kasus dan Implementasi Menggunakan Python*. 2021.

[12] J. Perintis Kemerdekaan Km, M. Syukri Mustafa, and I. Wayan Simpen, "PROSIDING SEMINAR ILMIAH SISTEM INFORMASI DAN TEKNOLOGI INFORMASI Pusat Penelitian dan Pengabdian Pada Masyarakat (P4M) STMIK Dipanegara Makassar Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba."

[13] I. A. Nikmatun and I. Waspada, "IMPLEMENTASI DATA MINING UNTUK KLASIFIKASI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR," *Jurnal SIMETRIS*, vol. 10, no. 2, 2019.

[14] R. K. Dinata, H. Akbar, and N. Hasdyna, "Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus," *ILKOM Jurnal Ilmiah*, vol. 12, no. 2, pp. 104–111, Aug. 2020, doi: 10.33096/ilkom.v12i2.539.104-111.

[15] N. M. Putry, "KOMPARASI ALGORITMA KNN DAN NAÏVE BAYES UNTUK KLASIFIKASI DIAGNOSIS PENYAKIT DIABETES MELLITUS," *EVOLUSI : Jurnal Sains dan Manajemen*, vol. 10, no. 1, Apr. 2022, doi: 10.31294/evolusi.v10i1.12514.

[16] N. Meilani and O. Nurdiawan, "Data Mining untuk Klasifikasi Penderita Kanker Payudara Menggunakan Algoritma K-Nearest Neighbor," 2023. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer.

[17] J. Supriyanto, D. Alita, and A. R. Isnain, "Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 4, no. 1, pp. 74–80, Mar. 2023, doi: 10.33365/jatika.v4i1.2468.