JURNAL DIPAKOMTI
Pusat Penelitian dan Pengabdian kepada Masyarakat (P3M) Universitas Dipa Makassar
Jl. Perintis Kemerdekaan Km. 9 Makassar

Analisis Dan Penerapan Algoritma Naive Bayes Untuk Klasifikasi Retensi Arsip Di Dinas Perhubungan Provinsi Sulawesi Selatan

Sisilia Gallaran, Nidia Matande*, Baharuddin Rahman, Sri Wahyuni
Jurusan Teknik Informatika, Universitas Dipa Makassar, Makassar Sulawesi Selatan
e-mail: [1]sisiliaglrn@gmail.com, [2]nidiamatande02@gmail.com,
[3]baharuddin.rahman@undipa.ac.id, [4]sriwahyuni@undipa.ac.id

Abstrak

Penelitian ini membahas permasalahan ketidakefisienan dan ketidakkonsistenan dalam proses klasifikasi retensi arsip di Dinas Perhubungan Provinsi Sulawesi Selatan yang berpotensi menyebabkan kesalahan penentuan masa simpan arsip dan keterlambatan layanan informasi. Data yang digunakan berasal dari 1330 arsip berbagai kegiatan administrasi, teknis, dan surat menyurat pada periode 2011–2023. Model klasifikasi dibangun menggunakan algoritma Naive Bayes dengan representasi teks TF-IDF, serta pembagian data 80% pelatihan dan 20% pengujian. Hasil penelitian menunjukkan kinerja model yang sangat baik dengan akurasi 99,62%, serta nilai precision dan recall di atas 0,99 untuk kedua kelas retensi, yaitu MUSNAH (arsip yang harus dimusnahkan) dan PERMANEN (arsip yang disimpan selamanya). Model ini dapat mengklasifikasikan arsip baru tanpa label secara otomatis, sehingga meningkatkan efisiensi, akurasi, dan mendukung digitalisasi pengelolaan arsip pemerintahan.

Kata kunci*—Naive Bayes, retensi arsip, klasifikasi, TF-IDF, manajemen arsip.*

Abstract

This research discusses the problems of inefficiency and inconsistency in the archive retention classification process at the Dinas Perhubungan Provinsi Sulawesi Selatan, which have the potential to cause errors in determining archive retention periods and delays in information services. The dataset consists of 1,330 archival records covering administrative, technical, and correspondence documents from 2011 to 2023. The classification model was developed using the Naive Bayes algorithm with TF-IDF text representation and an 80:20 train-test split. The model achieved outstanding performance with 99.62% accuracy, and precision and recall values above 0.99 for both retention classes: MUSNAH (records to be destroyed) and PERMANEN (records to be permanently preserved). The model can automatically classify new unlabeled archives, improving efficiency and accuracy while supporting the digitalization of archival management in government institutions.

Keywords—*Naive Bayes, archival retention, classification, TF-IDF, records management.*

1. PENDAHULUAN

Dinas Perhubungan Provinsi Sulawesi Selatan memiliki tanggung jawab besar dalam pengelolaan sistem transportasi dan infrastruktur di wilayahnya. Sebagai bagian dari tugasnya, dinas perhubungan mengelola berbagai arsip yang terkait dengan regulasi, administrasi, dan operasional transportasi. Pengelolaan arsip ini sangat penting untuk mendukung efisiensi

operasional dan kepatuhan terhadap peraturan yang berlaku. Di Dinas Perhubungan, arsip yang dimiliki mencakup berbagai jenis dokumen, mulai dari izin kendaraan, surat perizinan operasional, laporan inspeksi, hingga dokumen terkait proyek infrastruktur transportasi.

Menurut Kamus Besar Bahasa Indonesia (KBBI), secara umum retensi adalah penyimpanan atau penahanan. Retensi adalah mengurangi jumlah arsip dengan cara memisahkan arsip inaktif [1]. Menurut The Liang Gie, arsip adalah kumpulan catatan atau surat yang dapat digunakan kembali sehingga disusun dengan tujuan dan mempunyai pengaruh ketika catatan tersebut ditemukan kembali. Pencatatan merupakan cara yang paling umum digunakan untuk menyimpan dan mengumpulkan surat-surat atau catatan-catatan yang terdapat dalam suatu berkas sehingga laporan-laporan tersebut dapat ditemukan kembali jika diperlukan [2]. Pengelolaan arsip yang efektif sangat penting untuk memastikan informasi yang dibutuhkan dapat diakses dengan mudah dan cepat, serta mendukung pengambilan keputusan yang tepat waktu [3].

Di Dinas Perhubungan Provinsi Sulawesi Selatan, khususnya pada Sub- Bagian Umum, Kepegawaian, dan Hukum, arsip-arsip tersebut disimpan di ruangan arsip dengan menggunakan berbagai tempat penyimpanan, seperti lemari arsip, brankas, dan tempat penyimpanan lainnya. Meskipun sistem pengarsipan ini dapat dianggap cukup efisien dan aman, namun tidak sepenuhnya menghindarkan instansi dari potensi masalah yang bisa terjadi, seperti kerusakan pada tempat penyimpanan atau risiko bencana yang dapat merusak arsip secara fisik. Hal ini menjadi kekhawatiran, mengingat pentingnya keberlanjutan dan keselamatan arsip dalam menunjang kelancaran operasional administrasi. Pengalaman pribadi penulis yang pernah melaksanakan Praktek Kerja Lapangan (PKL) atau magang di instansi yang kebetulan ditempatkan di Sub-bagian Umum, Kepegawaian, dan Hukum yang terlibat langsung dengan pengarsipan tersebut turut mendukung pandangan ini, meskipun cukup tertata dengan baik, tetapi menghadapi tantangan terhadap faktor-faktor eksternal yang sulit diprediksi [4].

Beberapa penelitian sebelumnya telah membahas penerapan metode klasifikasi dalam bidang kearsipan, namun sebagian besar belum secara khusus meneliti penerapan algoritma Naive Bayes dalam konteks arsip pemerintahan daerah dengan karakteristik data yang kompleks dan beragam. Misalnya, penelitian [5] menerapkan Naive Bayes untuk klasifikasi retensi arsip dengan akurasi mencapai 95%, namun penelitian tersebut dilakukan di lingkungan pendidikan dan belum mencakup arsip pemerintahan daerah. Penelitian lain oleh Triayudi (2018) menggunakan klasifikasi naive bayes dan c4.5 untuk prediksi penerima bantuan jaminan kesehatan tidak menghasilkan model prediktif otomatis. Sementara itu, penelitian oleh Lisdiatini (2024) di Dinas Kearsipan Kota Madiun hanya menggunakan metode deskriptif kualitatif tanpa pengujian performa model klasifikasi. Berdasarkan hal tersebut, masih terdapat kesenjangan penelitian dalam penerapan algoritma Naive Bayes untuk klasifikasi otomatis arsip di lingkungan pemerintahan daerah dengan dataset yang lebih besar dan variatif [5], [6], [7].

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menerapkan algoritma Naive Bayes dalam klasifikasi retensi arsip pada Dinas Perhubungan Provinsi Sulawesi Selatan [8]. Penerapan ini diharapkan dapat meningkatkan efisiensi dan akurasi dalam pengelolaan arsip, serta mempercepat proses identifikasi kategori retensi arsip antara MUSNAH dan PERMANEN. Selain itu, penelitian ini juga diharapkan dapat memberikan kontribusi nyata terhadap digitalisasi sistem manajemen arsip di instansi pemerintahan, guna mendukung tata kelola dokumen yang lebih efektif, transparan, dan berbasis teknologi informasi [9], [10], [11].

Implementasi Naive Bayes dalam pengelolaan arsip di Dinas Perhubungan diharapkan dapat meningkatkan akurasi dalam menentukan masa simpan arsip serta meningkatkan efisiensi dalam proses klasifikasi arsip. Dengan sistem klasifikasi yang lebih otomatis dan tepat, diharapkan proses pengelolaan arsip dapat dilakukan dengan lebih cepat, sehingga arsip yang diperlukan dapat dengan mudah diakses oleh staf yang membutuhkan informasi tersebut. Selain itu, penerapan algoritma ini juga dapat mengurangi kesalahan manusia dan meningkatkan kualitas pengelolaan arsip secara keseluruhan, yang pada gilirannya akan meningkatkan kualitas layanan yang diberikan oleh Dinas Perhubungan Provinsi Sulawesi Selatan kepada masyarakat. Penerapan algoritma Naive Bayes dalam pengelolaan arsip dapat menghasilkan sistem yang lebih efisien dan

akurat, serta meminimalkan risiko kesalahan manusia dalam pengelolaan data [11], [12], [13], [14], [15].

2. METODE PENELITIAN

2.1 Jenis Penelitian

Jenis penelitian yang digunakan dalam studi ini adalah penelitian kuantitatif deskriptif. Penelitian kuantitatif berfokus pada pengumpulan dan analisis data numerik menggunakan metode statistik, sedangkan pendekatan deskriptif bertujuan untuk memberikan gambaran yang sistematis, faktual, dan akurat mengenai fenomena yang diteliti. Dalam konteks ini, penelitian ini bertujuan untuk mendeskripsikan dan menganalisis proses klasifikasi retensi arsip menggunakan algoritma Naive Bayes berdasarkan data arsip dari Dinas Perhubungan Provinsi Sulawesi Selatan. Melalui pendekatan ini, diharapkan diperoleh pemahaman yang komprehensif mengenai tingkat akurasi, efektivitas, serta potensi penerapan model klasifikasi dalam mendukung pengelolaan arsip secara digital dan efisien.

2.2 Tahapan Penelitian

Dalam penelitian ini, proses analisis klasifikasi menggunakan algoritma Naive Bayes dilakukan secara terstruktur dengan beberapa tahapan yang dirancang agar menghasilkan model yang akurat dan andal. Berikut adalah tahapan yang dilakukan dalam proses ini:

a. Pengumpulan Data

Pada tahap awal, data arsip yang digunakan dalam penelitian ini dikumpulkan dari periode tahun 2011 hingga 2023. Sumber data berasal dari dataset arsip yang telah tersedia pada sistem pengelolaan arsip di Dinas Perhubungan Provinsi Sulawesi Selatan. Dataset tersebut berisi kumpulan berkas arsip yang dihasilkan dari kegiatan administrasi dan operasional instansi yang telah berjalan. Data yang dikumpulkan mencakup informasi yang lengkap, relevan, dan sesuai dengan kebutuhan analisis klasifikasi retensi arsip, sehingga dapat digunakan sebagai dasar dalam proses pembentukan model klasifikasi.

b. Dataset

Data yang telah dikumpulkan diorganisasikan menjadi sebuah dataset terstruktur yang terdiri atas sejumlah fitur (variabel) dan label (kelas) yang digunakan dalam proses klasifikasi. Dataset ini bersumber dari arsip Dinas Perhubungan Provinsi Sulawesi Selatan yang mencakup berbagai jenis dokumen administrasi, teknis, dan surat menyurat yang dihasilkan selama periode 2011 hingga 2023.Secara keseluruhan, dataset berjumlah 1.330 arsip, dengan dua kelas utama pada variabel target, yaitu MUSNAH dan PERMANEN. Kelas MUSNAH merepresentasikan arsip yang tidak lagi memiliki nilai guna dan dapat dimusnahkan sesuai jadwal retensi, sedangkan kelas PERMANEN menunjukkan arsip yang memiliki nilai historis, hukum, atau administratif jangka panjang dan wajib disimpan selamanya. Adapun fitur-fitur utama dalam dataset mencakup: Nomor Berkas, Judul Berkas, Nomor Item, Klasifikasi Arsip, Uraian Informasi Arsip, Kurun Waktu, Tingkat Perkembangan, Jumlah, dan Keterangan. Dari keseluruhan atribut tersebut, kolom Uraian Informasi Arsip dan Judul Berkas menjadi fokus utama dalam proses klasifikasi berbasis teks karena mengandung konteks semantik yang relevan terhadap penentuan retensi arsip.

c. Preprocessing Data

Data yang belum siap digunakan seringkali memiliki berbagai masalah, seperti data kosong (missing values), anomali, atau skala yang tidak konsisten. Oleh karena itu, dilakukan preprocessing untuk membersihkan dan mempersiapkan data. Tahapan ini meliputi:

1) Penanganan data hilang (imputasi nilai hilang)
2) Normalisasi data untuk mengatur skala nilai.
3) Encoding data kategoris jika diperlukan.

d. Evaluasi Kesiapan Data

Setelah seluruh tahapan preprocessing selesai dilakukan, langkah selanjutnya adalah melakukan evaluasi kesiapan data untuk memastikan bahwa dataset telah memenuhi syarat

untuk digunakan dalam pelatihan model. Evaluasi ini mencakup pengecekan terhadap kelengkapan data, keseimbangan antar kelas (MUSNAH dan PERMANEN), serta kesesuaian format data hasil transformasi TF-IDF. Apabila masih ditemukan inkonsistensi atau ketidaksesuaian, maka proses dikembalikan ke tahap preprocessing untuk dilakukan penyesuaian ulang. Namun, jika data telah dinyatakan valid dan siap digunakan, maka proses dilanjutkan ke tahap pembentukan model.

e. Pembentukan Model

Pada tahap ini dilakukan pembangunan model klasifikasi menggunakan algoritma Naive Bayes. Data yang telah siap dibagi menjadi dua bagian, yaitu 80% sebagai data latih (training data) dan 20% sebagai data uji (testing data). Pembagian dilakukan menggunakan metode Stratified Split, agar proporsi kelas MUSNAH dan PERMANEN tetap seimbang di kedua subset data. Algoritma Naive Bayes kemudian dilatih menggunakan data latih untuk menghitung probabilitas prior dan posterior dari setiap kelas berdasarkan distribusi kata yang dihasilkan melalui TF-IDF. Hasil dari tahap ini adalah model klasifikasi yang mampu memprediksi kelas retensi arsip pada data baru secara otomatis.

f. Validasi Hasil Evaluasi

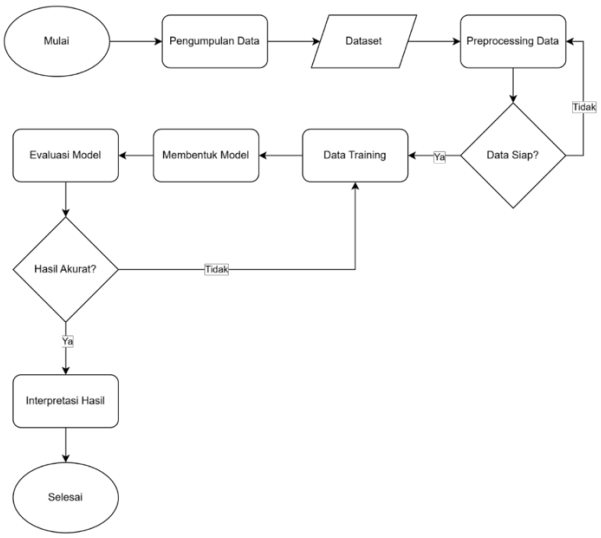
Model yang telah dibentuk dievaluasi menggunakan data testing. Evaluasi dilakukan untuk mengukur performa model berdasarkan metrik tertentu, seperti akurasi, presisi, recal, dan F1-Score. Hal ini bertujuan untuk menilai keandalan model dalam melakukan klasifikasi.

g. Evaluasi Model

Tahap validasi dilakukan untuk memastikan konsistensi dan keandalan hasil evaluasi model. Jika hasil pengujian menunjukkan nilai akurasi dan metrik lainnya belum memenuhi kriteria yang diharapkan, maka dilakukan penyesuaian parameter dan pelatihan ulang model (retraining). Namun, jika hasil evaluasi menunjukkan performa yang memadai dan stabil, maka model dinyatakan valid dan siap untuk diinterpretasikan.

h. Interpretasi Hasil

Tahap ini merupakan analisis dan penafsiran terhadap kinerja model yang telah divalidasi. Hasil dari metrik evaluasi dijelaskan secara kuantitatif untuk menunjukkan sejauh mana algoritma Naive Bayes mampu mengklasifikasikan arsip dengan benar. Selain itu, dilakukan juga peninjauan terhadap pola-pola kata yang berpengaruh besar dalam menentukan kelas MUSNAH dan PERMANEN. Interpretasi ini memberikan gambaran mengenai kemampuan model untuk diterapkan secara nyata dalam sistem retensi arsip, serta kontribusinya dalam meningkatkan efisiensi dan akurasi pengelolaan arsip digital.



Gambar 1. Tahapan Penelitian

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Data

Data yang digunakan dalam penelitian ini berasal dari arsip Dinas Perhubungan Provinsi Sulawesi Selatan. Data mencakup berkas arsip dari berbagai jenis kegiatan administrasi, teknis, dan surat menyurat yang terjadi selama periode 2011 hingga 2023. Dataset yang digunakan terdiri dari beberapa atribut utama sebagai berikut:

Tabel 1 Deskripsi Data

Nama Kolom	Keterangan
Jenis_Arsip	Deskripsi atau nama kegiatan/berkas arsip
Status_Perkembangan	Status perkembangan fisik arsip
Kurun_waktu	Tahun arsip dibuat
Jumlah	Jumlah unit arsip
Bentuk	Bentuk fisik arsip (BERKAS atau TEKSTUAL)
Nasib_akhir	Label kategori retensi arsip (target label)

3.2 Preproccesing Data

Pada penelitian ini, preprocessing dilakukan secara bertahap untuk memastikan bahwa data bersih, konsisten, dan siap digunakan dalam proses pelatihan model klasifikasi. Berikut ini adalah tahapan-tahapan preprocessing yang diterapkan:

a. Penanganan duplikasi data

Dalam penelitian ini, fungsi drop_duplicates() dari pustaka pandas digunakan untuk menghapus baris-baris yang identik secara keseluruhan.

```
Jumlah total data sebelum pengecekan duplikat: 1342
Jumlah baris duplikat yang ditemukan: 12

Contoh data duplikat:
                                        Jenis_Arsip Tingkat_Perkembangan  \
13   PENGAWASAN SERTA PEMBINAAN ATAS LALU LINTAS AN...                 COPY
198  SK KENAIKAN PANGKAT PENGATUR ATASA NAMA WAHYUD...                 COPY
446                                    SKP IRFAN                     COPY
450                              SKP ABD. RACHMAN AZIZ                 COPY
467  REKOMENDASI NOMOR ; 0013/REKOM-MPU /SDPU/VI/20...                 ASLI

    Kurun_Waktu Jumlah   Bentuk Nasib_Akhir
13         2002      1   BERKAS      MUSNAH
198        2010      1  BERKAS       MUSNAH
446        2017      9  BERKAS       MUSNAH
450        2017      6  BERKAS       MUSNAH
467        2017      9   BERKAS    PERMANEN

Jumlah data setelah duplikat dihapus: 1330
```

Gambar 2. Hasil Preproccesing Data

Gambar 2 merupakan hasil preproccesing data, dimana ditemukan jumlah data yang duplikat sebanyak 12 data, sehingga dilakukan penghapusan data yang duplikat, sehingga setelah pengahpusan data yang terhindar dari duplikasi sebanyak 1330 data

b. Pembersihan teks

Setelah melakukan penanganan pada duplikasi data dan data kosong, tahapan selanjutnya adalah melakukan pembersihan pada teks dengan menggunakan metode lowercasing, penghapusan tanda baca, stopword removal, dan steaming. Hal ini penting untuk memastikan bahwa data yang digunakan benar-benar bersih, berikut merupakan hasil pembersihan teks yang telah dikakukan:

```
                                        Jenis_Arsip  \
1  DAFTAR NAMA PNS YANG BELUM DIAMBIL SUMPAH JANJ...
2  SURAT KEPUTUSAN MENTERI PERHUBUNGAN TENTANG KE...
3  KEPUTUSAN MENTERI PERHUBUNGAN NOMOR KP.100/6/K...
4  KEPUTUSAN GUBERNUR SULAWESI SELATAN TENTANG PE...
5                            IJAZAH SMA A.N FITRIAH

                                    Preprocessed_Teks
1  daftar nama pns ambil sumpah janji pns dinas h...
2  surat putus menteri hubung naik pangkat regule...
3  putus menteri hubung nomor kpkpphb angkat calo...
4  putus gubernur sulawesi selatan angkat calon p...
5                           ijazah sma an fitriah
```

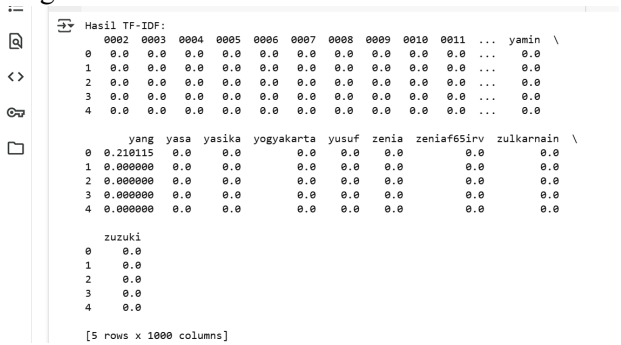
Gambar 3. Hasil Pembersihan Teks

Gambar 3 merupakan hasil pembersihan teks, dalam hal ini dilakukan konversi dari huruf capital menjadi lowercase, tujuannya adalah untuk memudahkan proses analisis dan

menyelaraskan semua huruf dan kata yang ada, karena huruf dan kata yang menjadi titik analisis dalam penelitian ini.

c. TF-IDF Vectorization

Setelah tahapan pembersihan teks, tahapan selanjutnya adalah melakukan transformasi teks menjadi representasi numerik menggunakan metode TF-IDF (Term Frequency – Inverse Document Frequency). TF-IDF digunakan karena lebih efektif daripada Count Vectorizer dalam menilai kepentingan kata dalam dokumen.



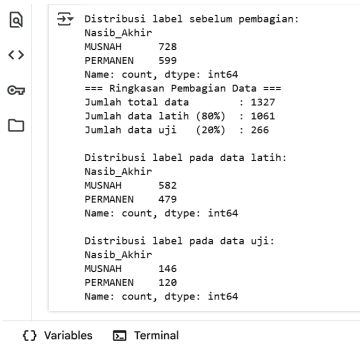
```
Hasil TF-IDF:
   0002  0003  0004  0005  0006  0007  0008  0009  0010  0011  ...  yamin  \
0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0  ...    0.0
1   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0  ...    0.0
2   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0  ...    0.0
3   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0  ...    0.0
4   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0  ...    0.0

       yang  yasa  yasika  yogyakarta  yusuf  zenia  zeniaf65irv  zulkarnain  \
0  0.210115   0.0     0.0         0.0    0.0    0.0          0.0         0.0
1  0.000000   0.0     0.0         0.0    0.0    0.0          0.0         0.0
2  0.000000   0.0     0.0         0.0    0.0    0.0          0.0         0.0
3  0.000000   0.0     0.0         0.0    0.0    0.0          0.0         0.0
4  0.000000   0.0     0.0         0.0    0.0    0.0          0.0         0.0

   zuzuki
0     0.0
1     0.0
2     0.0
3     0.0
4     0.0

[5 rows x 1000 columns]
```

Gambar 4. Hasil TF-IDF Vectorization

Gambar 4 merupakan hasil konversi setiap kata yang telah dibersihkan menjadi bentuk numerik, hal ini dilakukan untuk mengurutkan tingkat kepentingan dan seberapa banyak kata tersebut muncul dalam arsip, hal ini dilakukan untuk mempermudah proses analisis.

d. Split Data

Setelah tahapan preprocessing selesai dilakukan, data kemudian dibagi ke dalam dua bagian utama menggunakan metode train-test split yang disediakan oleh library Scikit-Learn. Metode ini membagi dataset secara acak dengan proporsi tertentu, dalam penelitian ini 80% data digunakan sebagai data latih (training set) dan 20% data digunakan sebagai data uji (testing set).



```
Distribusi label sebelum pembagian:
Nasib_Akhir
MUSNAH      728
PERMANEN    599
Name: count, dtype: int64
=== Ringkasan Pembagian Data ===
Jumlah total data       : 1327
Jumlah data latih (80%)  : 1061
Jumlah data uji   (20%)  : 266

Distribusi label pada data latih:
Nasib_Akhir
MUSNAH      582
PERMANEN    479
Name: count, dtype: int64

Distribusi label pada data uji:
Nasib_Akhir
MUSNAH      146
PERMANEN    120
Name: count, dtype: int64
```

Variables Terminal

Gambar 5. Hasil Pembagian Data Uji dan Data Latih

Gambar 5 merupakan hasil pembagian antara data uji dan latih, dimana data latih yang digunakan sebanyak 1061 dan data uji sebanyak 266 data, selain itu diketahui juga bahwa terdapat 582 dari data latih yang digunakan memilikin kelas musnah dan 479 memiliki kelas permanen, hal ini penting untuk memastikan bahwa persebaran data yang digunakan untuk data latih merata dan menghasilkan analisis yang akurat.

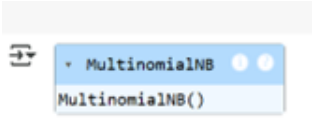
3.3 Pembentukan Model Naïve Bayes

Pada tahap ini dilakukan proses pelatihan model klasifikasi menggunakan algoritma Naive Bayes. Algoritma yang digunakan adalah Multinomial Naive Bayes karena sesuai untuk menangani data teks yang telah dikonversi menjadi bentuk numerik melalui metode TF-IDF. Fitur utama yang digunakan dalam pembentukan model adalah hasil ekstraksi dari kolom Jenis_Arsip yang telah diolah melalui proses TF-IDF Vectorization.

Model Naive Bayes dibentuk menggunakan pustaka scikit-learn dengan implementasi default dari MultinomialNB. Proses pelatihan dilakukan dengan menggunakan data latih (X_train, y_train) untuk mempelajari pola hubungan antara istilah-istilah dalam deskripsi arsip dan status

retensinya. Model yang dihasilkan bersifat generatif dan mampu memberikan probabilitas terhadap masing-masing kelas berdasarkan fitur yang diberikan.

Setelah proses pelatihan selesai, model disimpan dalam format digital menggunakan pustaka joblib agar dapat digunakan kembali untuk klasifikasi data arsip baru.



Gambar 6. Hasil Pembentukan Model Naïve Bayes

Gambar 6 merupakan hasil pembentukan model Naïve Bayes, model yang telah dilatih disimpan dalam bentuk mulnomial(NB), fungsi ini nantinya yang akan dipanggil saat ingin melakukan klasifikasi terhadap arsip yang belum memiliki label.

3.4 Evaluasi Kinerja Model

Setelah model Naive Bayes berhasil dibentuk, langkah selanjutnya adalah melakukan evaluasi terhadap performanya dalam melakukan klasifikasi data arsip. Evaluasi dilakukan menggunakan data uji sebesar 20% dari total dataset, yang sebelumnya telah dipisahkan secara stratifikasi untuk memastikan distribusi kelas tetap proporsional. Beberapa metrik evaluasi yang digunakan antara lain:

a. Accuracy

Metrik ini mengukur persentase prediksi yang benar dibandingkan dengan total data uji. Akurasi memberikan gambaran umum performa model, namun dapat menyesatkan jika data tidak seimbang.

b. Precision

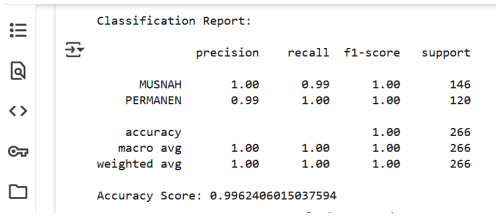
Presisi menunjukkan proporsi prediksi benar terhadap semua prediksi yang dibuat untuk suatu kelas. Presisi tinggi menandakan bahwa kesalahan tipe I (false positive) relatif kecil.

c. Recal

Recal mengukur proporsi data positif yang berhasil terdeteksi oleh model. Nilai ini penting untuk memastikan bahwa arsip yang seharusnya masuk kategori tertentu tidak terlewatkan.

d. F1-Score

F1-Score adalah rata-rata harmonik dari presisi dan Recal, memberikan metrik seimbang ketika distribusi kelas tidak merata.

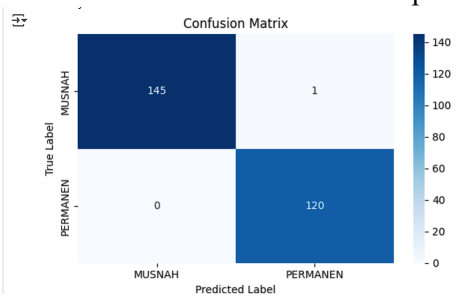


Classification Report:

	precision	recall	f1-score	support
MUSNAH	1.00	0.99	1.00	146
PERMANEN	0.99	1.00	1.00	120
accuracy			1.00	266
macro avg	1.00	1.00	1.00	266
weighted avg	1.00	1.00	1.00	266

Accuracy Score: 0.9962406015037594

Gambar 7. Clasification Report



Gambar 8. Hasil Evaluasi Model Naïve Bayes

Dari hasil evaluasi pada gambar 7, diketahui bahwa model menunjukkan kinerja yang sangat baik. Tingkat akurasi mencapai 99,62%, yang berarti hampir seluruh prediksi terhadap data uji sesuai dengan label sebenarnya. Untuk metrik presisi, nilai yang dicapai adalah 1.00 untuk

kelas MUSNAH dan 0.99 untuk PERMANEN, menunjukkan bahwa prediksi positif yang dilakukan oleh model sangat akurat.

Sementara itu, nilai Recal menunjukkan 0.99 untuk kelas MUSNAH dan 1.00 untuk PERMANEN. Hal ini mengindikasikan bahwa model mampu mengidentifikasi sebagian besar data dari masing-masing kelas dengan benar. F1-Score yang tinggi, yaitu 1.00 untuk kedua kelas, mencerminkan keseimbangan yang baik antara presisi dan Recal.

Hasil dari confusion matrix pada gambar 8 juga memperkuat evaluasi ini. Dari 266 data uji, hampir seluruhnya diprediksi dengan tepat, dan hanya terdapat sedikit kesalahan klasifikasi. Capaian ini menunjukkan bahwa model Naive Bayes tidak hanya kuat secara statistik, tetapi juga konsisten dan dapat diandalkan untuk proses klasifikasi otomatis arsip.

Model klasifikasi retensi arsip yang dibangun menggunakan algoritma Naive Bayes berhasil menunjukkan performa yang sangat tinggi berdasarkan hasil evaluasi terhadap data uji. Dari total 266 data uji, model mampu memprediksi 265 data secara tepat, dengan nilai akurasi mencapai 99,62%. Hanya satu data yang diklasifikasikan secara keliru. Nilai precision dan recal masing-masing berada pada rentang 0,99 hingga 1,00, baik untuk kelas MUSNAH maupun PERMANEN, dengan nilai rata-rata macro F1-Score mencapai 1,00. Dari distribusi kelas dalam data uji, terdapat 146 data berlabel MUSNAH dan 120 data berlabel PERMANEN. Meskipun distribusinya tidak sepenuhnya seimbang, model mampu menangani kedua kelas dengan akurasi dan stabilitas prediksi yang seimbang. Hal ini menunjukkan bahwa fitur Jenis_Arsip yang telah melalui proses transformasi teks (TF-IDF) berhasil menangkap pola linguistik yang membedakan kedua kelas tersebut secara signifikan.

3.5 Pengujian Model

Sebagai bentuk validasi eksternal, dilakukan pengujian terhadap data arsip baru sebanyak 1000 entri yang tidak memiliki label Nasib Akhir. Data ini diambil dari arsip tahun yang berbeda dan berisi deskripsi kegiatan yang belum pernah dilihat oleh model sebelumnya. Berikut merupakan hasil pengujian terhadap arsip yang tidak memiiki label:

```
✅ Prediksi selesai! File hasil disimpan sebagai 'hasil_prediksi.xlsx'.
```

	Jenis_Arsip	Tingkat_Perkembangan	Kurun_Waktu	Jumlah	Bentuk	Prediksi_Nasib_Akhir
0	Kumpulan keputusan Direktur Jenderal Perhubung...	1993	Asli	1	Kertas	MUSNAH
1	Surat Perjanjian SPK (Nomor : 550/010/SPK/IV/1...	2012	Asli	1	Kertas	MUSNAH
2	Laporan Hasil Pelaksanaan Tugas Pengadaan Peng...	2006	Asli	1	Kertas	MUSNAH
3	Kumpulan Undangan Sekretaris Daerah Provinsi S...	2018	Asli	1	Kertas	MUSNAH
4	Surat- surat Undangan asistensi usulan tingkat...	2017	Asli	1	Kertas	MUSNAH
...	...	...	...	...	...	...
995	Kumpulan Berkas. Kartu Pengawasan	1999	Asli	1	Kertas	MUSNAH
996	Laporan bulan Pertama. Pengawasan Pengadaan da...	2017	Asli	1	Kertas	MUSNAH
997	Surat Perintah Kerja (SPK) Tentang Pekerjaan...	2007	Copy	1	Kertas	MUSNAH
998	Hasil Pelaksanaan Kegiatan Pengawasan dan Pend...	2018	Copy	1	Kertas	MUSNAH
999	Surat Perintah Kerja (SPK). Pekerjaan Pengaw...	2007	Copy	1	Kertas	MUSNAH

1000 rows × 6 columns

Gambar 9. Hasil Uji Coba Model

Gambar 9 merupakan hasil uji coba model dengan menggunaakan 1000 data arsip tanpa label, tujuan uji coba untuk mengetahui apakah model sudah mampu untuk memberikan label otomatis terhadap arsip. Berdasarkan hasil uji coba, diketahui bahwa model mampu memberikan label kepada 1000 arsip, sehingga dapat disimpulkan bahwa model sudah mampu bekerja untuk melabeli arsip yang ada.

4. KESIMPULAN

Penerapan algoritma Naive Bayes terbukti mampu meningkatkan akurasi hingga 99,62% dengan precision dan recall di atas 0,99 pada kedua kelas (MUSNAH dan PERMANEN). Analisis teks pada kolom Jenis_Arsip memperlihatkan bahwa model dapat mengenali pola kata yang khas untuk tiap kelas. Implementasi metode ini di Dinas Perhubungan Provinsi Sulawesi Selatan dapat

menjadi solusi nyata untuk otomatisasi klasifikasi arsip, mempercepat pengambilan keputusan retensi, serta mendukung digitalisasi pengelolaan arsip pemerintahan.

5. SARAN

Berdasarkan hasil penelitian dan potensi penerapan model klasifikasi retensi arsip, beberapa saran pengembangan lebih lanjut adalah:

a. Pengayaan Dataset: Perlu menambah variasi jenis arsip, periode waktu, dan deskripsi agar model lebih adaptif.
b. Pengembangan Model: Eksplorasi algoritma lain seperti Random Forest, SVM, atau deep learning untuk perbandingan performa.
c. Integrasi Aplikasi: Model diimplementasikan ke sistem berbasis web/desktop dengan antarmuka sederhana agar mudah digunakan.
d. Penerapan Luas: Jika terbukti efektif, sistem dapat diperluas ke instansi lain guna mendukung transformasi digital pemerintahan.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Dipa Makassar dan semua pihak yang telah memberikan dukungan dan bantuan selama penelitian ini berlangsung. Penghargaan khusus disampaikan kepada Bapak Dr. Baharuddin Rahman, M.Hum dan Ibu Sri Wahyuni, S.Kom., MT atas bimbingan dan inspirasinya dalam penyusunan penelitian ini. Terima kasih juga kepada orang tua yang selalu memberikan doa, dukungan, dan semangat tanpa henti dalam setiap langkah penulis. Serta Semua pihak yang telah membantu dengan cara apa pun.

DAFTAR PUSTAKA

[1] W. Via Trisna and T. Purnama Sari, “Accompanying the implementation of medical records of the retention system in the Annisa Pekanbaru Hospital in 2023,” ARSY: Aplikasi Riset kepada Masyarakat, vol. 4, no. 2, 2024. [Online]. Available: http://journal.al-matani.com/index.php/arsy
[2] Laboratory Office Bandung State Polytechnic, “Bisnis dan Digital (JIMaKeBiDi),” vol. 1, no. 2.
[3] A. Wicaksana et al., “Penerapan sistem informasi pengelolaan arsip untuk meningkatkan efisiensi administrasi,” Jurnal Teknologi dan Informasi, vol. 10, no. 1, pp. 12–25, 2022.
[4] M. Amirulloh et al., “Tantangan pengelolaan arsip secara manual pada Dinas Perhubungan dan solusi teknologi yang dapat diterapkan,” Jurnal Administrasi Publik, vol. 11, no. 3, pp. 45–56, 2023.
[5] B. Harijanto, Y. Ariyanto, and L. Miftahurroifa, “Penerapan algoritma Naive Bayes untuk klasifikasi retensi arsip,” n.d.
[6] A. Triayudi and R. T. Aldisa, “Analisis performa algoritma klasifikasi Naive Bayes dan C4.5 untuk prediksi penerima bantuan jaminan kesehatan,” Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), vol. 7, no. 2, pp. 262–269, 2023. [Online]. Available: https://doi.org/10.35870/jtik.v7i2.756
[7] F. A. Faizah et al., “Analisis performa algoritma klasifikasi Naive Bayes dan C4.5 untuk prediksi penerima bantuan jaminan kesehatan,” Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), vol. 7, no. 2, pp. 262–269, 2023. [Online]. Available: https://doi.org/10.35870/jtik.v7i2.756
[8] A. Saputra et al., “Analisis efektivitas algoritma Naive Bayes dalam pengelolaan arsip digital,” Journal of Computer Science and Information Technology, vol. 8, no. 2, pp. 102–115, 2023.
[9] S. Kusumadewi and S. Hartati, “Penerapan algoritma Naive Bayes untuk klasifikasi teks,”

Jurnal Teknologi dan Sistem Komputer, vol. 1, no. 1, pp. 1–8, 2013.
[10] B. Harijanto, Y. Ariyanto, and L. Miftahurroifa, "Penerapan algoritma Naive Bayes untuk klasifikasi retensi arsip," Jurnal Informatika Polinema, vol. 4, no. 2, p. 155, 2018. [Online]. Available: https://doi.org/10.33795/jip.v4i2.15
[11] H. Judul, F. T. Industri, and U. I. Indonesia, "Analisis sentimen X terhadap keamanan data menggunakan algoritma Naive Bayes classifier: Studi kasus pusat data nasional sementara (PDNS)," 2024.
[12] M. Rosyid Ash-Shodiq and D. Prasetia Caraka, "The effect of organized archiving on employee performance at the Business Attamami," 2024.
[13] R. Rokhmah and E. Nuryani, "Metode klasifikasi dalam machine learning," Jurnal Ilmiah Komputer, vol. 5, no. 2, pp. 45–56, 2020.
[14] A. Rakhman and A. Nugroho, "Analisis penggunaan Python dalam pengembangan aplikasi machine learning," Jurnal Teknologi Informasi dan Komunikasi, vol. 3, no. 1, pp. 23–30, 2021.
[15] R. O. Felani, H. Syaputra, and W. Cholil, "Analisa prilaku pengguna e-learning menggunakan algoritma K-means clustering," Jusikom: Jurnal Sistem Komputer Musirawas, n.d.