

**PERBANDINGAN METODE KLASIFIKASI ALGORITMA *NAÏVE*  
*BAYES* TANPA DAN DENGAN *KERNEL DENSITY ESTIMATION*  
(STUDI KASUS DATA *SELF DECLARE* BPJPH 2022)**

**SKRIPSI**



**AGUS HERMAWAN**

**H051191008**

**PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS HASANUDDIN  
MAKASSAR  
2023**

**PERBANDINGAN METODE KLASIFIKASI ALGORITMA *NAÏVE*  
*BAYES* TANPA DAN DENGAN *KERNEL DENSITY ESTIMATION*  
(STUDI KASUS DATA *SELF DECLARE* BPJPH 2022)**

**SKRIPSI**

Diajukan sebagai salah satu syarat memperoleh gelar Sarjana Sains  
Pada Program Studi Statistika Departemen Statistika  
Fakultas Matematika dan Ilmu Pengetahuan Alam  
Universitas Hasanuddin

**AGUS HERMAWAN  
H051191008**

**PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS HASANUDDIN  
MAKASSAR  
MEI 2023**

**LEMBAR PERNYATAAN KEOTENTIKAN**

Saya yang bertanda tangan di bawah ini menyatakan dengan sungguh-sungguh bahwa skripsi yang saya buat dengan judul:

**PERBANDINGAN METODE KLASIFIKASI ALGORITMA *NAÏVE BAYES* TANPA DAN DENGAN *KERNEL DENSITY ESTIMATION* (STUDI KASUS DATA *SELF DECLARE* BPJPH 2022)**

adalah benar hasil karya saya sendiri, bukan hasil plagiat dan belum pernah dipublikasikan dalam bentuk apapun.

Makassar, 31 Mei 2023




**AGUS HERMAWAN**  
H051191008

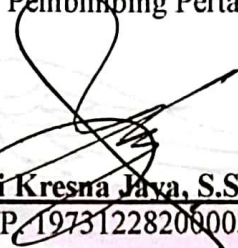
**PERBANDINGAN METODE KLASIFIKASI ALGORITMA *NAÏVE*  
BAYES TANPA DAN DENGAN *KERNEL DENSITY ESTIMATION*  
(STUDI KASUS DATA *SELF DECLARE* BPJPH 2022)**

Disetujui oleh:

Pembimbing Utama

  
Siswanto, S.Si., M.Si.  
NIP. 199201072019031012

Pembimbing Pertama

  
Andi Kresna Jaya, S.Si., M.Si.  
NIP. 197312282000031001

Ketua Program Studi

  
  
Dr. Arif Islamiyati, S.Si., M.Si.  
NIP. 197708082005012002

Pada 31 Mei 2023

## HALAMAN PENGESAHAN

Skripsi ini diajukan oleh:

Nama : Agus Hermawan  
NIM : H051191008  
Program Studi : Statistika  
Judul skripsi : Perbandingan Metode Klasifikasi Algoritma *Naïve Bayes*  
Tanpa Dan Dengan *Kernel Density Estimation* (Studi Kasus  
Data *Self Declare BPJPH 2022*)

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

### DEWAN PENGUJI

1. Ketua : Siswanto, S.Si., M.Si. (.....)
2. Sekretaris : Andi Kresna Jaya, S.Si., M.Si. (.....)
3. Anggota : Drs. Raupong, M.Si. (.....)
4. Anggota : Dra. Nasrah Sirajang, M.Si. (.....)

Ditetapkan di : Makassar

Tanggal : 31 Mei 2023

## KATA PENGANTAR

Segala puji senantiasa dipanjatkan kepada kehadiran Ilahirabbi Yang Mahakuasa yang telah memberikan kekuatan dan rahmat-Nya sehingga penyusunan skripsi ini dapat terselesaikan dengan baik. Skripsi dengan judul “Perbandingan Metode Klasifikasi Algoritma *Naïve Bayes* Tanpa Dan Dengan *Kernel Density Estimation* (Studi Kasus Data *Self Declare* BPJPH 2022)” ini merupakan salah satu rangkaian syarat akademik yang harus dipenuhi untuk memperoleh Gelar Sarjana Sains pada Program Studi Statistika, Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin.

Skripsi ini merupakan penelitian yang bertujuan untuk membandingkan persentase performa metode klasifikasi Algoritma *Naïve Bayes* tanpa dan dengan *Kernel Density Estimation* menggunakan data *Self Declare* BPJPH 2022 dan sebagai upaya dalam proses mengoptimalkan verifikasi dan validasi data sertifikasi halal dengan memanfaatkan teknologi komputasi data. Metode penelitian ini melibatkan algoritma *Naïve Bayes* dan optimasi *Kernel Density Estimation* yang kemudian diharapkan dapat memberikan wawasan dan referensi kepada BPJPH dalam klasifikasi data.

Penulis menyadari bahwa penyelesaian skripsi ini tidak terlepas dari bantuan dan dorongan yang diberikan oleh berbagai pihak yang secara konsisten memberikan bantuan baik secara moril maupun materil. Meskipun penulis memiliki keterbatasan dalam kemampuan dan pengetahuan, namun berkat bantuan dan dukungan tersebut, penulis berhasil menyelesaikan skripsi ini. Oleh karena itu, penulis ingin mengucapkan terima kasih yang sebesar-besarnya dan penghargaan yang tulus kepada semua pihak yang terlibat. Oleh karena itu, dengan penuh kesadaran dan kerendahan hati, pada kesempatan ini perkenankanlah penulis menyampaikan ucapan terima kasih dan penghargaan setinggi-tingginya kepada yang terhormat:

1. Terima kasih dan apresiasi yang setinggi-tingginya kepada diri sendiri yang telah berusaha dengan gigih dan tekun selama proses penyelesaian skripsi ini. Tidak mudah untuk melewati tantangan dan rintangan yang dihadapi selama

proses penyusunan, namun penulis membuktikan keberhasilannya dengan terselesaikannya skripsi ini.

2. Terima kasih yang tak terhingga kepada kedua orang tua tercinta atas dukungan dan pengorbanan yang tidak bisa penulis ungkapkan dengan kata-kata terhadap proses penyelesaian skripsi ini.
3. Terima kasih kepada Prof. Dr. Ir. Jamaluddin Jompa, M.Sc., selaku Rektor Universitas Hasanuddin beserta seluruh staf jajarannya.
4. Terima kasih kepada Bapak Dr. Eng. Amiruddin, selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin beserta seluruh staf jajarannya.
5. Terima kasih kepada Ibu Dr. Nurtiti Sunusi S.Si., M.Si., yang waktu itu masih menjabat sebagai Ketua Departemen Statistika atas dorongan dan motivasinya karena hampir setiap kali berpapasan selalu ditanya “kapan Agus naik ujian?”. Sekali lagi terimakasih banyak, Ibu. Saat ini penulis telah menyelesaikan apa yang selalu Ibu tanyakan dengan baik.
6. Terima kasih kepada Pak Siswanto, S.Si., M.Si., sebagai pembimbing utama dan penasehat akademik yang telah bersedia untuk meluangkan waktu untuk senantiasa menerima kendala penyusunan dan memberikan solusi yang terbaik.
7. Terima kasih kepada Pak Andi Kresna Jaya, S.Si., M.Si. sebagai pembimbing pertama yang selalu tampil dengan ciri khas pembawaannya yang bercanda santai dengan kalimat legendnya ketika ditanyakan sesuatu “saya tidak tahu, saya juga orang baru di sini” namun pada akhirnya tetap diberikan arahan dengan baik hingga tuntas. Sekali lagi terimakasih banyak, Pak.
8. Terima kasih kepada Pak Drs. Raupong, M.Si. dan Dra. Nasrah Sirajang, M.Si. sebagai dosen penguji yang telah bersedia meluangkan waktunya untuk memberikan penilaian dan masukan terhadap skripsi ini.
9. Terima kasih kepada segenap jajaran dosen matakuliah dan staf Departemen Statistika yang telah banyak membantu, memberikan ilmu-ilmunya, serta berbagai kemudahan lainnya yang diberikan selama menempuh pendidikan sarjana di Departemen Statistika.

10. Terima kasih kepada Badan Penyelenggara Jaminan Produk Halal (BPJPH) Kemenag RI yang telah memberikan izin penggunaan data untuk keperluan penelitian ini.
11. Terima kasih kepada pemilik NIM 1192050109 yang oleh penulis dijadikan sebagai motivator untuk bisa menuntaskan skripsi ini dengan baik.
12. Terima kasih kepada teman-teman dekat di Istana Tercinta yang setia menemani penulis dari masa kuliah perdana sampai akhir perkuliahan. Terima kasih kanda-kanda: Arief Rahman Nur yang selalu mau direpotkan atas jasa transportasinya, Fadilah Amirul Adhel, Muhammad Syamsul Bahri, Muhammad Ferdiansyah, dan Sapriadi Rasyid.
13. Terima kasih kepada Muhammad Fathurahman telah menjadi teman terbaik bagi penulis di masa perkuliahan. Terima kasih, Kanda.
14. Terima kasih kepada semua teman-teman di angkatan Statistika 2019 yang telah menerima penulis yang merupakan orang pendatang di Makassar.
15. Terima kasih yang setinggi-tingginya kepada seluruh pihak yang mungkin tidak sempat penulis sebutkan satu persatu. Terima kasih atas segala dukungan, partisipasi, dan apresiasinya yang diberikan kepada penulis.

Penulis juga menyadari bahwa skripsi ini masih jauh dari kata sempurna, namun ini hasil terbaik yang dapat diberikan oleh penulis pada penelitian ini. Oleh karena dengan segala kerendahan hati penulis mengucapkan permohonan maaf yang sebesar-besarnya. Akhir kata, semoga tulisan ini dapat memberikan manfaat untuk berbagai pihak.

Makassar, 31 Mei 2023  
Penulis

  
**AGUS HERMAWAN**  
NIM. H051191008



**PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR  
UNTUK KEPENTINGAN AKADEMIK**

---

Sebagai civitas akademik Universitas Hasanuddin, saya yang bertanda tangan di bawah ini:

Nama : Agus Hermawan  
NIM : H051191008  
Program Studi : Statistika  
Departemen : Statistika  
Fakultas : Matematika dan Ilmu Pengetahuan Alam  
Jenis Karya : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin **Hak Bebas Royalti Non-eksklusif (*Non-exclusive Royalty- Free Right*)** atas tugas akhir saya yang berjudul:

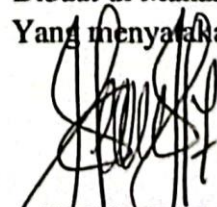
“Perbandingan Metode Klasifikasi Algoritma *Naïve Bayes* Tanpa Dan Dengan *Kernel Density Estimation* (Studi Kasus Data *Self Declare* BPJPH 2022)”

Beserta perangkat yang ada (jika diperlukan). Terkait dengan hal di atas, maka pihak universitas berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (database), merawat dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di Makassar, 31 Mei 2023

Yang menyatakan,

  
**AGUS HERMAWAN**  
NIM. H051191008

**ABSTRAK**

Sertifikasi halal memberikan jaminan kehalalan produk kepada konsumen Muslim di seluruh dunia. Dengan adanya sertifikasi halal, konsumen Muslim dapat dengan percaya diri mengonsumsi produk tersebut. BPJPH sebagai auditor resmi di Indonesia, memiliki cukup banyak kebutuhan dalam upaya proses verifikasi dan validasi data pengajuan sertifikasi halal. Melalui pendekatan *Data Science*, teknologi kemampuan *machine learning* diperlukan untuk mengoptimalkan proses verifikasi dan validasi data tersebut. Tujuan dari penelitian ini adalah untuk memberikan solusi dari permasalahan tersebut melalui metode klasifikasi *Naïve Bayes*. Metode klasifikasi *Naïve Bayes* perlu diterapkan metode optimasi seperti *Kernel Density Estimation* (KDE) untuk mendapatkan hasil klasifikasi yang lebih baik. Sehingga, metode klasifikasi *Naïve Bayes* tanpa optimasi dan dengan optimasi ini perlu dibandingkan untuk memastikan kinerja algoritma *Naïve Bayes* KDE dapat menjawab kekurangan dari *Naïve Bayes* dan sejauh mana dapat menangani data kompleks dan tidak berdistribusi normal. Hasil dari penelitian ini adalah mendapatkan nilai persentase performa model klasifikasi *Naïve Bayes* tanpa optimasi: akurasi 87,6%, *recall* 85,4%, presisi 88,8%, dan  $F_{measure}$  87,1%. Sementara itu, persentase performa model klasifikasi *Naïve Bayes* KDE mencapai: akurasi 97,5%, *recall* 95,9%, presisi 98,9%, dan  $F_{measure}$  97,8%. Sehingga, dapat disimpulkan bahwa algoritma klasifikasi *Naïve Bayes* KDE menunjukkan peningkatan persentase performa klasifikasi yang cukup baik dibandingkan dengan model *Naïve Bayes* tanpa optimasi dengan peningkatan sebesar 9,9%.

**Kata Kunci:** *Naïve Bayes*, *Kernel Density Estimation*, Klasifikasi, Sertifikasi Halal, *Self Declare*

**ABSTRACT**

*Halal certification provides assurance of the halal status of products to Muslim consumers worldwide. With the presence of halal certification, Muslim consumers can confidently consume the products. BPJPH, as the official auditor in Indonesia, has significant needs in the process of verification and validation of halal certification application data. Through a Data Science approach, the machine learning technology is required to optimize the verification and validation process of the data. The aim of this research is to provide a solution to this issue through the Naïve Bayes classification method. The Naïve Bayes classification method needs to be implemented with optimization methods such as Kernel Density Estimation (KDE) to obtain better classification results. Therefore, the Naïve Bayes classification method without optimization and with optimization needs to be compared to ensure that the percentage performance of the Naïve Bayes KDE algorithm can address the limitations of Naïve Bayes and to what extent it can handle complex and non-normally distributed data. The results of this research showed that the percentage performance of the Naïve Bayes classification model without optimization achieved an accuracy of 87.6%, recall of 85.4%, precision of 88.8%, and  $F_{measure}$  of 87.1%. Meanwhile, the percentage performance of the Naïve Bayes KDE classification model reached an accuracy of 97.5%, recall of 95.9%, precision of 98.9%, and  $F_{measure}$  of 97.8%. Thus, it can be concluded that the Naïve Bayes KDE classification algorithm demonstrates a significant improvement in classification percentage performance compared to the Naïve Bayes model without optimization, with an increase of 9.9%.*

**Keywords:** *Naïve Bayes, Kernel Density Estimation, Clasification, Halal Certification, Self Declare*

## DAFTAR ISI

<b>HALAMAN SAMPUL</b> .....	<b>i</b>
<b>HALAMAN JUDUL</b> .....	<b>ii</b>
<b>LEMBAR PERNYATAAN KEOTENTIKAN</b> .....	<b>iii</b>
<b>HALAMAN PENGESAHAN</b> .....	<b>v</b>
<b>KATA PENGANTAR</b> .....	<b>vi</b>
<b>HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI</b> .....	<b>ix</b>
<b>ABSTRAK</b> .....	<b>x</b>
<b>ABSTRACT</b> .....	<b>xi</b>
<b>DAFTAR ISI</b> .....	<b>xii</b>
<b>DAFTAR GAMBAR</b> .....	<b>xiv</b>
<b>DAFTAR TABEL</b> .....	<b>xv</b>
<b>DAFTAR LAMPIRAN</b> .....	<b>xvi</b>
<b>BAB I PENDAHULUAN</b> .....	<b>1</b>
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah .....	3
1.4 Tujuan Penelitian .....	3
1.5 Manfaat Penelitian .....	3
<b>BAB II TINJAUAN PUSTAKA</b> .....	<b>4</b>
2.1 Sertifikasi Halal di Indonesia .....	4
2.2 <i>Self Declare</i> .....	4
2.3 Pra-pemrosesan Data .....	5
2.3.1 <i>Tokenizing</i> .....	5
2.3.2 <i>Case Folding</i> .....	6
2.3.3 <i>Stopword Removal</i> .....	6
2.4 Ekstraksi Fitur dan Normalisasi.....	6
2.4.1 <i>Term Frequency</i> .....	6
2.4.2 <i>Inverse Document Frequency</i> .....	7
2.4.3 <i>Term Frequency_Inverse Document Frequency</i> .....	7
2.5 Klasifikasi Dokumen .....	8

2.5.1 Data Latih.....	8
2.5.2 Data Uji .....	9
2.6 <i>Naïve Bayes</i> .....	9
2.7 <i>Kernel Density Estimation</i> .....	11
2.8 Uji Persentase performa.....	13
<b>BAB III METODOLOGI PENELITIAN .....</b>	<b>15</b>
3.1 Data.....	15
3.2 Metode Analisis Data .....	16
<b>BAB IV HASIL DAN PEMBAHASAN .....</b>	<b>17</b>
4.1 Deskripsi Data .....	17
4.2 Pra-pemrosesan Data .....	17
4.2.1 <i>Punctuation Removal</i> .....	17
4.2.2 <i>Case Folding</i> .....	18
4.2.3 <i>Stopwords Removal</i> .....	18
4.2.4 <i>Tokenizing</i> .....	19
4.3 Visualisasi Teks dengan <i>Word Cloud</i> .....	19
4.4 <i>Count Vectorizer</i> .....	20
4.5 Pembagian Data Latih dan Data Uji .....	21
4.6 <i>TF_IDF Vectorizer</i> pada Data Latih .....	21
4.7 Klasifikasi <i>Naïve Bayes</i> .....	25
4.8 Uji Persentase performa Klasifikasi <i>Naïve Bayes</i> .....	26
4.9 Ekstraksi Fitur dengan KDE.....	29
4.10 Klasifikasi <i>Naïve Bayes</i> KDE.....	34
4.11 Uji Persentase performa Klasifikasi <i>Naïve Bayes</i> KDE .....	35
4.12 Perbandingan Uji Persentase performa.....	38
<b>BAB V PENUTUP .....</b>	<b>39</b>
5.1 Kesimpulan .....	39
5.2 Saran .....	40
<b>DAFTAR PUSTAKA .....</b>	<b>41</b>
<b>LAMPIRAN.....</b>	<b>44</b>

**DAFTAR GAMBAR**

**Gambar 4.1** Visualisasi teks dengan *Word Cloud* ..... 20

**Gambar 4.2** *Confusion Matrix* hasil prediksi *Naïve Bayes*..... 27

**Gambar 4.3** Akurasi Model *Naïve Bayes* ..... 28

**Gambar 4.4** *Confusion Matrix* hasil prediksi *Naïve Bayes* KDE ..... 36

**Gambar 4.5** Akurasi Model *Naïve Bayes* KDE ..... 37

**Gambar 4.6** Perbandingan Uji Persentase performa..... 38

## DAFTAR TABEL

<b>Tabel 2.1</b> <i>Confusion Matrix</i> .....	13
<b>Tabel 3.1</b> Sampel data .....	15
<b>Tabel 4.1</b> Dataset .....	17
<b>Tabel 4.2</b> Hasil tahapan <i>punctuation removal</i> .....	18
<b>Tabel 4.3</b> Hasil tahapan <i>case folding</i> .....	18
<b>Tabel 4.4</b> Hasil tahapan <i>stopwords removal</i> .....	19
<b>Tabel 4.5</b> Hasil tahapan <i>tokenizing</i> .....	19
<b>Tabel 4.6</b> <i>Count vectorizer</i> dataset .....	20
<b>Tabel 4.7</b> Jumlah kemunculan fitur dataset.....	21
<b>Tabel 4.8</b> Proporsi pembagian data latih dan data uji .....	21
<b>Tabel 4.9</b> <i>Count vectorizer</i> data latih .....	22
<b>Tabel 4.10</b> Jumlah kemunculan fitur dengan selisih minimal 30.....	22
<b>Tabel 4.11</b> Hasil <i>Term Frequency_Invers Document Frequency</i> Positif .....	23
<b>Tabel 4.12</b> Hasil <i>Term Frequency_Invers Document Frequency</i> Negatif .....	25
<b>Tabel 4.13</b> Nilai peluang pada contoh data uji .....	25
<b>Tabel 4.14</b> Hasil prediksi <i>Naïve Bayes</i> .....	26
<b>Tabel 4.15</b> Fitur pada data latih.....	29
<b>Tabel 4.16</b> Fitur dengan jumlah selisih minimal 30 .....	30
<b>Tabel 4.17</b> Penduga $\hat{x}$ dan $\hat{x}_i$ Kelas Positif .....	30
<b>Tabel 4.18</b> Penduga $\hat{x}$ dan $\hat{x}_i$ Kelas Negatif .....	31
<b>Tabel 4.19</b> Bobot KDE pada Kelas Positif.....	32
<b>Tabel 4.20</b> Bobot KDE pada Kelas Negatif .....	34
<b>Tabel 4.21</b> Nilai peluang pada contoh data uji fitur KDE.....	34
<b>Tabel 4.22</b> Hasil prediksi <i>Naïve Bayes</i> KDE.....	35

## DAFTAR LAMPIRAN

<b>Lampiran 1</b> <i>Count vectorizer</i> jumlah kemunculan fitur pada dataset .....	45
<b>Lampiran 2</b> <i>Count vectorizer</i> jumlah kemunculan fitur pada data latih.....	46
<b>Lampiran 3</b> Jumlah kemunculan fitur pada masing-masing kelas .....	47
<b>Lampiran 4</b> Bobot TF-IDF Kelas Positif.....	48
<b>Lampiran 5</b> Bobot TF-IDF Kelas Negatif .....	49
<b>Lampiran 6</b> Bobot KDE Kelas Positif.....	50
<b>Lampiran 7</b> Bobot KDE Kelas Negatif .....	51
<b>Lampiran 8</b> Hasil klasifikasi data uji <i>Naïve Bayes</i> .....	52
<b>Lampiran 9</b> Hasil klasifikasi data uji <i>Naïve Bayes</i> KDE .....	54



## BAB I PENDAHULUAN

### 1.1 Latar Belakang

Indonesia merupakan negara dengan mayoritas penduduk yang beragama Islam dengan jumlah penduduk sebanyak 272,23 juta jiwa pada Juni 2021. Berdasarkan data dari Dukcapil Kementerian Dalam Negeri, sebanyak 86,88% dari total penduduk Indonesia beragama Islam, diikuti oleh Kristen, Katolik, Hindu, Buddha, Konghucu, dan menganut aliran kepercayaan. Hal ini menjadikan status kehalalan produk sangat sensitif di Indonesia. Pemerintah membentuk Badan Penyelenggara Jaminan Produk Halal (BPJPH) dengan MUI sebagai auditor produk yang didaftarkan. BPJPH membuka jalur pendaftaran Sertifikasi Halal Reguler (berbayar) dan Sertifikasi Halal Gratis (tidak berbayar) dengan skema *self declare* sebagai jalur pernyataan halal yang melibatkan Pendamping Proses Produk Halal (PPH) yang telah tersertifikasi (Fajrin & Mohammad, 2021).

Sejauh ini, BPJPH melakukan kurasi data secara manual dengan verifikasi dan validasi satu persatu oleh karyawan. Namun, dengan pertumbuhan jumlah pendaftaran yang akan terus meningkat hingga mencapai jutaan data, proses kurasi manual akan menjadi tidak efektif. Oleh karena itu, perlu adanya pemanfaatan teknologi komputasi data seperti model *machine learning* dalam ilmu statistika untuk mengoptimalkan proses verifikasi dan validasi data.

Penelitian sebelumnya tentang Klasifikasi Interaksi Kampanye di Media Sosial dari data *Twitter* dengan menggunakan *Naïve Bayes Kernel Estimator* (Nugroho dkk., 2019) yang memiliki nilai akurasi sebesar 80,14%. Penelitian ini menggunakan berfokus pada data pengajuan jalur Sertifikasi Halal Gratis (tidak berbayar). Algoritma yang dipilih adalah *Naïve Bayes* karena mampu melakukan pengklasifikasian dengan metode probabilitas dan statistik yang dapat memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Saleh dkk., 2015). *Naïve Bayes* memiliki keuntungan, menurut Manalu dkk., (2017), yaitu hanya memerlukan jumlah data pelatihan yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Selain itu, menurut

Hidayanti dkk., (2020), *Naïve Bayes* sering kali mampu bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks daripada yang diharapkan.

Algoritma *Naïve Bayes* memiliki kelemahan saat melakukan proses *training* dengan *dataset* yang memiliki bobot *atribute* yang tidak seragam, sehingga persentase performa klasifikasinya kurang baik (Rezki dkk., 2020). Selain itu, Metode *Naïve Bayes* memiliki asumsi kuat tentang independensi antar fitur yang digunakan dalam proses klasifikasi. Namun, pada data nama produk *self declare* BPJPH 2022 ini tentunya ada ketergantungan atau hubungan yang kompleks di setiap kata yang menjadi penyusun sebuah nama produk. Solusi untuk mengatasi hal ini adalah dengan menggunakan fungsi *kernel density* (Silverman, 2018a).

*Kernel Density Estimation* (KDE) merupakan teknik statistik untuk memperkirakan fungsi densitas probabilitas dari sekelompok data dengan menemukan kernel terbaik (Pérez dkk., 2009). Penggunaan KDE pada data teks dengan jumlah data yang besar merupakan teknik yang tepat untuk digunakan pada masalah ini dan kebutuhan di BPJPH. Algoritma *Naïve Bayes Classifier Kernel Density Estimation* merupakan hasil optimalisasi dari penggunaan KDE pada algoritma *Naïve Bayes*. Sehingga, kedua metode klasifikasi ini hasilnya perlu dibandingkan untuk memastikan kinerja algoritma *Naïve Bayes* KDE dapat menjawab kekurangan dari *Naïve Bayes* dan sejauh mana dapat menangani data kompleks dan tidak berdistribusi normal. Oleh karena itu, penelitian ini berjudul "Perbandingan Metode Klasifikasi Algoritma *Naïve Bayes* Tanpa Dan Dengan *Kernel Density Estimation* (Studi Kasus Data *Self Declare* BPJPH 2022)".

## 1.2 Rumusan Masalah

Berdasarkan uraian latar belakang, maka rumusan masalah yang akan diangkat pada penelitian ini yaitu: Bagaimana perbandingan persentase performa metode klasifikasi Algoritma *Naïve Bayes* tanpa dan dengan *Kernel Density Estimation* pada studi kasus data *self declare* BPJPH 2022?

### 1.3 Batasan Masalah

Batasan masalah yang digunakan pada penelitian ini adalah:

1. Data yang digunakan adalah data pengajuan sertifikasi halal tidak berbayar pada data *self declare* BPJPH 2022.
2. Optimasi menggunakan metode *Kernel Density Estimation* dengan fungsi kernel *Gaussian* dan *Bandwidth* menggunakan *Silvermen Rule of Thumb*.

### 1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Mendapatkan hasil perbandingan persentase performa metode klasifikasi Algoritma *Naïve Bayes* tanpa dan dengan *Kernel Density Estimation* (studi kasus data *self declare* BPJPH 2022).

### 1.5 Manfaat Penelitian

Manfaat yang diperoleh dari hasil penelitian ini adalah sebagai berikut:

1. Mendapatkan kinerja penggunaan algoritma *Naïve Bayes* dan *Naïve Bayes Kernel Density Estimation* dalam sebuah klasifikasi data.
2. Referensi kepada instansi terkait tentang hasil klasifikasi dengan menggunakan algoritma *Naïve Bayes* dan *Naïve Bayes Kernel Density Estimation*.

## BAB II TINJAUAN PUSTAKA

### 2.1 Sertifikasi Halal di Indonesia

Spesifikasi sertifikasi halal di Indonesia terbatas pada bahan-bahan yang berasal dari hewan, tumbuhan, mikroba, dan bahan yang dihasilkan melalui proses kimiawi, proses biologi, atau proses rekayasa genetik dalam proses produksi pembuatan produk halal baik itu nantinya dijadikan sebagai bahan baku, bahan olahan, bahan tambahan, maupun bahan penolong harus halal menurut syariat agama (Fajrin & Mohammad, 2021).

### 2.2 *Self Declare*

*Self declare* merupakan program Sertifikasi Halal Gratis yang diberikan oleh pemerintah untuk percepatan capaian target 10 juta produk tersertifikasi halal. Kriteria produk yang masuk kategori *self declare* terdapat pada Surat Keputusan Kepala BPJPH Nomor 33 tahun 2022 tentang Kriteria *self declare*, yaitu: 1) Produk tidak berisiko atau menggunakan bahan yang sudah dipastikan kehalalannya. 2) Proses produksi yang dipastikan kehalalannya dan sederhana, 3) Memiliki hasil penjualan tahunan (omset) maksimal 500 juta. 4) Memiliki Nomor Induk Berusaha (NIB), 5) Memiliki lokasi, tempat, dan alat proses produksi (PPH) yang terpisah dengan lokasi, tempat dan alat proses produk tidak halal. 6) Memiliki atau tidak memiliki surat izin edar (PIRT/MD/UMOT/UKOT), Sertifikat Laik Higiene Sanitasi (SLHS) untuk produk makanan/minuman dengan daya simpan kurang dari 7 hari, atau izin industri lainnya atas produk yang dihasilkan dari dinas/instansi terkait. 7) Memiliki outlet dan/atau fasilitas produksi paling banyak 1 lokasi. 8) Secara aktif telah memproduksi 1 tahun sebelum permohonan sertifikasi halal. 9) Produk yang dihasilkan berupa barang (bukan jasa atau usaha retoran, kantin, catering, dan kedai/rumah/warung makan). 10) Bahan yang digunakan sudah dipastikan kehalalannya dibuktikan dengan sertifikat halal atau termasuk dalam daftar bahan nabati murni dari alam atau positif lis. 11) Tidak menggunakan bahan yang berbahaya. 12) Telah diverifikasi kehalalannya oleh Pendamping Proses Produk Halal (PPPH). 13) jenis produk/kelompok produk yang disertifikasi halal tidak mengandung unsur hewan hasil sembelihan, kecuali berasal dari produsen

atau rumah potong hewan yang sudah bersertifikat halal. 14) Menggunakan peralatan produksi dengan teknologi sederhana atau dilakukan secara manual dan/atau semi manual. 15) Proses pengawetan produk yang dihasilkan tidak menggunakan teknik radiasi, rekayasa genetika, penggunaan ozon, dan kombinasi metode pengawetan. 16) Melengkapi dokumen pengajuan sertifikasi halal dengan mekanisme pernyataan pelaku usaha secara online melalui SIHALAL (Kemenag, 2022).

Jenis produk yang masuk kategori *self declare* terdapat pada Surat Keputusan Kepala BPJPH Nomor 33 tahun 2022 tentang Kriteria *self declare*, yaitu: 1) Susu dan analognya, 2) Lemak, minyak, dan emulsi minyak. 3) Es untuk dimakan termasuk sherbet dan sortbet. 4) Buah dan sayur dengan pengolahan dan penambahan bahan tambahan pangan. 5) Kembang gula/permen dan cokelat. 6) Serealias dan produk sereal yang merupakan produk turunan dari biji sereal, akar, dan umbi kacang-kacangan. 7) Produk bakeri. 8) Ikan dan produk perikanan. 9) Telur olahan dan produk telur hasil olahan. 10) Gula dan pemanis termasuk madu. 11) Garam, rempah, sup, saus, salad, serta produk protein. 12) Makanan ringan siap santap. 13) Kelompok bahan lainnya. 14) Sari buah dan sari sayuran. 15) Konsentrat sari buah dan sari sayuran. 16) Minuman berbasis air. 17) Kopi. 18) Minuman biji-bijian dan sereal panas. 19) Minuman berbasis susu. 20) Minuman tradisional. 21) Jamu. 22) Ekstrak herbal terstandar. 23) Ekstrak bahan alam (Kemenag, 2022).

### **2.3 Pra-pemrosesan Data**

Pra-pemrosesan Data (Data Preprocessing) merupakan sebuah tahapan paling awal dalam pemrosesan data. Pada tahapan ini, data yang tidak terstruktur akan diubah menjadi data yang terstruktur yang umumnya dilakukan pada data dalam bentuk teks dan disebut *text preprocessing*. *Text preprocessing* merupakan tindakan menghilangkan karakter-karakter tertentu yang terkandung dalam dokumen, seperti: titik, koma, tanda petik dan lain-lain, serta mengubah semua huruf kapital menjadi huruf kecil (Lestari, 2019).

#### **2.3.1 Tokenizing**

*Tokenizing* adalah proses pembentukan kata dari urutan karakter dalam sebuah dokumen. Dalam sistem awal, “kata” didefinisikan sebagai urutan karakter

alfanumerik dengan panjang 3 atau lebih banyak, diakhiri oleh spasi atau karakter khusus lainnya, tetapi bergantung pada ekstraksi informasi dan transformasi *query* untuk menangani masalah yang sulit. Di banyak kasus, aturan tambahan ditambahkan ke *tokenizer* untuk menangani beberapa yang khusus karakter, untuk memastikan bahwa token dan token dokumen akan cocok. Dalam *tokenizing*, setiap kalimat dalam dokumen diubah menjadi token (kata-kata *string*) (Lestari, 2019).

### 2.3.2 Case Folding

Proses *case folding* adalah proses mengubah seluruh huruf menjadi huruf kecil. Pada proses ini karakter-karakter ‘A’-‘Z’ yang terdapat pada data diubah kedalam karakter ‘a’-‘z’. Karakter-karakter selain huruf ‘a’ sampai ‘z’ (tanda baca dan angka-angka) akan dihilangkan dari data dan dianggap sebagai *delimiter*. *Delimiter* adalah urutan satu atau lebih karakter yang digunakan untuk menentukan batas pemisah (Jumeilah, 2017).

### 2.3.3 Stopword Removal

Bahasa manusia diisi dengan kata-kata fungsi yaitu kata-kata yang memiliki sedikit makna terpisah dari kata-kata lain. Kata-kata ini adalah bagian dari bagaimana kita mendeskripsikan kata benda dalam teks, dan mengekspresikannya konsep seperti lokasi atau kuantitas. Preposisi, seperti seperti “yang”, “dari”, “di”, dan sebagainya (Lestari, 2019).

## 2.4 Ekstraksi Fitur dan Normalisasi

Ekstraksi fitur didasarkan pada metode *Term Frequency-Inverse Document Frequency* (TF-IDF) standar yang merupakan salah satu metode populer yang digunakan dalam beberapa domain teks. TF-IDF bekerja dengan memilih kata-kata signifikan dan memberikan bobot tinggi kedalam frekuensi tinggi didalam dokumen yang berbeda, tetapi relatif jarang di seluruh korpus. Tahap ini setiap kata (*term*) hasil *preprocessing* dari seluruh dokumen yang akan digunakan untuk perhitungan *cluster weight* dihitung bobotnya. Perhitungan bobot yang digunakan adalah TF, IDF, dan TF-IDF (Lestari, 2019).

### 2.4.1 Term Frequency

*Term Frequency* merupakan salah satu metode untuk menghitung bobot setiap kata dalam teks berdasarkan frekuensi kemunculan kata. Prinsip pada

metode ini adalah bahwa setiap kata memiliki nilai kepentingan sebanding dengan jumlah kemunculan kata tersebut dalam teks. Pada *Term Frequency* (TF) formula yang dapat digunakan adalah TF logaritmik, hal ini untuk menghindari dominansi dokumen yang mengandung sedikit *term* dalam query, namun mempunyai frekuensi yang tinggi. Berikut ini adalah persamaan yang dipakai untuk menghitung bobot TF:

$$TF = \begin{cases} 1 + \log(f_{t,d}), & f_{t,d} > 0 \\ 0, & f_{t,d} = 0 \end{cases} \quad (2.1)$$

untuk nilai  $f_{t,d}$  adalah frekuensi *term* (t) pada *document* (d).

#### 2.4.2 *Inverse Document Frequency*

Sebuah file terdiri dari satu atau lebih dokumen dan setiap dokumen memiliki *term* atau kata yang dapat kita hitung jumlahnya. *Document Frequency* (DF) merupakan tahapan paling mendasar dalam proses pembobotan. Cara menghitung DF hanya melihat dokumen mana yang mengandung kata yang dicari. Perhitungan DF dilakukan dengan cara menghitung jumlah dokumen dimana sebuah fitur itu muncul. Dalam seleksi fitur, metode DF adalah kriteria yang paling sederhana dan mudah untuk dataset yang biasanya fokus pada kemunculan kata pada keseluruhan koleksi teks. Keunikan dari metode ini adalah kata yang jarang muncul pada keseluruhan koleksi kata dinilai lebih berharga. *Inverse Document Frequency* (IDF) dihitung dengan menggunakan formula sebagai berikut:

$$w_{x,y} = tf_{x,y} \log \frac{N}{f_{t,d}} \quad (2.2)$$

Keterangan:

$w_{x,y}$  : bobot dari kata

$tf_{x,y}$  : frekuensi suatu *term*

N : jumlah semua dokumen dalam korpus

$f_{t,d}$  : frekuensi *term* (t) pada *document* (d)

#### 2.4.3 *Term Frequency\_Inverse Document Frequency*

TF dan IDF untuk menghitung bobot *term* menghasilkan persentase performa yang lebih baik. Kombinasi bobot dari sebuah kata *t* pada teks *d* didefinisikan dalam persamaan berikut

$$TF\_IDF_t = w_t = TF \times IDF \quad (2.3)$$

Keterangan:

$TF\_IDF_t$  : bobot dari *term* ( $t$ )

$w_t$  : bobot dari *term* ( $t$ )

$TF$  : frekuensi *term* ( $t$ ) pada dokumen

$IDF$  : *Inverse Document Frequency*

## 2.5 Klasifikasi Dokumen

Klasifikasi dokumen adalah suatu proses pengelompokan dokumen sesuai dengan pembahasan di dalamnya. Klasifikasi dokumen merupakan masalah yang mendasar namun sangat penting karena manfaatnya cukup besar mengingat jumlah dokumen yang ada setiap waktu semakin bertambah. Sebuah dokumen dapat dikelompokkan ke dalam kategori tertentu berdasarkan kata-kata dan kalimat-kalimat yang ada di dalam dokumen tersebut. Kata atau kalimat yang terdapat di dalam sebuah dokumen memiliki makna tertentu dan dapat digunakan sebagai dasar untuk menentukan kategori dari suatu dokumen lainnya.

Klasifikasi data terdiri dari dua langkah proses. Pertama adalah *learning* (fase latihan), dimana algoritma klasifikasi dibuat untuk menganalisa data latihan kemudian direpresentasikan dalam bentuk *rule* atau model klasifikasi. Proses kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari *rule* atau model klasifikasi. Untuk dapat melakukan klasifikasi otomatis ini perlu adanya penerapan *machine learning* dengan algoritma terbaik berdasarkan karakteristik datanya.

### 2.5.1 Data Latih

Data latih adalah sebuah data yang dilatih untuk menemukan sebuah model yang cocok untuk digunakan dalam proses pengklasifikasian (Han dkk., 2018). Umumnya data-data dari dokumen yang telah dilatih sebelumnya akan menggunakan metode *supervised learning*, *Supervised learning* adalah pendekatan *machine learning* yang menggunakan data-data yang sudah diberi label atau dataset-nya sudah diketahui jenis klasifikasinya (Sen dkk., 2020). Sehingga, Data-data yang telah dirancang ini diharapkan melatih “*supervise*” algoritma untuk klasifikasi ataupun prediksi sebuah kasus dengan akurat (Kashif dkk., 2021).



### 2.5.2 Data Uji

Data uji adalah data yang digunakan untuk mengevaluasi persentase performa model yang telah dilatih (Plante & Werner, 2018). Setelah model dibangun dengan menggunakan data latih, model tersebut diterapkan pada data uji untuk mengetahui seberapa baik model tersebut dapat mengatasi masalah atau mengenali pola dalam data baru (Costa & Sánchez, 2022). Data uji harus berbeda dari data latih, karena model harus dapat bekerja dengan baik pada data yang tidak pernah dilihat sebelumnya (Almasan dkk., 2022). Jika data uji sama dengan data latih, maka tidak mungkin untuk mengetahui seberapa baik model dapat mengatasi masalah atau mengenali pola dalam data baru (Mahmood dkk., 2021).

### 2.6 Naïve Bayes

Algoritma *Naïve Bayes* didasarkan pada Teorema *Bayes* yang menjelaskan bagaimana mengubah probabilitas *priori* menjadi probabilitas *posteriori* (Senika dkk., 2022). Pada *text mining*, *Naïve Bayes* digunakan untuk mengklasifikasikan dokumen ke dalam kelas-kelas tertentu (Faradhillah, 2016). Algoritma ini mengasumsikan bahwa setiap fitur (kata) dalam dokumen independen satu sama lain, yang dikenal sebagai asumsi *Naïve* (Maulana & Yahya, 2019). Ini berarti bahwa probabilitas kemunculan sebuah kata dalam dokumen tidak dipengaruhi oleh kemunculan kata lain dalam dokumen tersebut (Zainab dkk., 2020).

*Naïve Bayes* memiliki beberapa keuntungan seperti proses *training* yang cepat, sederhana dan membutuhkan data latih yang sedikit (Pintoko & Lhaksana, 2018). Namun, asumsi *naïve* yang digunakan dalam algoritma ini dapat membuat hasil klasifikasi kurang akurat jika fitur-fitur dalam dokumen terkait satu sama lain. Peluang bersyarat terjadinya  $C$  dengan syarat  $X$  telah terjadi dapat dirumuskan sebagai berikut (Delizo dkk., 2020).

$$P(C|X) = \frac{P(X \cap C)}{P(X)}, P(X) > 0$$

Keterangan:

$P(C|X)$  : Peluang kejadian  $C$  dengan syarat  $X$  telah terjadi

$P(X)$  : Peluang kejadian  $X$

$P(X \cap C)$  : Peluang kejadian  $X$  dengan syarat  $C$  telah terjadi

Untuk menentukan probabilitas fitur  $x_i$  pada kelas  $C$ , dituliskan sebagai berikut:

$$\begin{aligned}
 P(C|X_1) &= \frac{P(X_1 \cap C)}{P(X_1)} \\
 P(C|X_2) &= \frac{P(X_2 \cap C)}{P(X_2)} \\
 &\vdots \\
 P(C|X_n) &= \frac{P(X_n \cap C)}{P(X_n)}
 \end{aligned}$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor-faktor syarat yang mempengaruhi nilai peluang yang hampir mustahil untuk dianalisis satu persatu. Kompleksnya faktor-faktor syarat yang mempengaruhi nilai peluang menyebabkan perhitungan tersebut menjadi sulit untuk dilakukan, maka digunakan asumsi independensi yang sangat tinggi (*naïve*), bahwa masing-masing petunjuk  $(X_1, X_2, X_3, \dots, X_n)$  saling bebas (*independent*) satu sama lain. Sehingga, berlaku suatu kesamaan sebagai berikut.

$$\begin{aligned}
 P(X_a|X_b) &= \frac{P(X_a \cap X_b)}{P(X_b)} & (2.4) \\
 &= \frac{P(X_a) \times P(X_b)}{P(X_b)} \\
 &= P(X_a)
 \end{aligned}$$

Berdasarkan Persamaan (2.4), syarat peluang menjadi lebih sederhana. Sehingga, perhitungan menjadi mungkin dilakukan. Selanjutnya,  $P(C|X)$  menggunakan aturan perkalian dapat diuraikan menjadi:

$$\begin{aligned}
 P(C|X) &\propto P(X \cap C) \\
 &= P(X_1 \cap X_2 \cap X_3 \cap \dots \cap X_n \cap C) \\
 &= P(X_1|X_2 \cap X_3 \cap \dots \cap X_n \cap C) \times P(X_2 \cap X_3 \cap \dots \cap X_n \cap C) \\
 &= P(X_1|X_2 \cap X_3 \cap \dots \cap X_n \cap C) \\
 &\quad \times P(X_2|X_3 \cap X_4 \cap \dots \cap X_n \cap C) \dots (X_{n-1}|X_n \cap \dots \cap C) \\
 &\quad \times P(X_n|C).P(C)
 \end{aligned}$$

Untuk menentukan peluang fitur  $x_i$  pada kelas  $C$  adalah sebagai berikut:

$$\begin{aligned}
 P(C|X_1, X_2, \dots, X_n) &\propto P(C)P(X_1|C)P(X_2|C) \dots P(X_n|C) & (2.5) \\
 &= P(C) \times \prod_{i=1}^n P(X_i|C)
 \end{aligned}$$

Persamaan ini juga disebut sebagai persamaan *posterior* yang didapatkan dari hasil kali antara nilai *prior* dan *likelihood* (Watraton dkk., 2020).

## 2.7 Kernel Density Estimation

*Kernel Density Estimation* (KDE) adalah metode statistik non-paramaterik yang digunakan untuk mendapatkan estimasi dari fungsi densitas dan menaksir distribusi probabilitas dari data (Haben & Giasemidis, 2016). KDE dapat digunakan sebagai optimasi dari *Naïve Bayes*, yang merupakan algoritma klasifikasi yang populer digunakan dalam *text mining* (Ji dkk., 2019). *Naïve Bayes* menganggap bahwa semua kata dalam dokumen independen satu sama lain, sementara KDE mengambil korelasi antara kata-kata dalam dokumen menjadi perhitungan. Dengan menggunakan KDE, model *Naïve Bayes* diperkuat dan dapat memberikan hasil yang lebih akurat (Nugroho dkk., 2019).

KDE digunakan dalam proses *feature extraction* (Ekstraksi fitur) dari data teks sebelum algoritma *Naïve Bayes* diterapkan (Qin & Xiao, 2018). *Feature extraction* dilakukan dengan mengekstrak kata-kata atau frasa dari sebuah dokumen atau korpus (Sethuraman & Athisayam, 2021). KDE digunakan untuk menaksir distribusi probabilitas dari sebuah variabel acak berdasarkan data yang diambil dari variabel tersebut yang kemudian dalam *text mining* variabel-variabel tersebut dinamakan fitur (Węglarczyk, 2018).

Setelah *feature extraction*, *Naïve Bayes* diterapkan pada data yang telah terekstraksi (Zhang dkk., 2018). Algoritma ini akan menyesuaikan pembobotan dan klasifikasi dari data latih yang digunakan untuk mengklasifikasikan data baru (Bullmann dkk., 2018). Secara umum, KDE memberikan optimasi pada algoritma *Naïve Bayes* dengan memberikan estimasi yang lebih baik dari probabilitas kata dalam dokumen dengan cara mengambil korelasi antara kata-kata dalam dokumen menjadi perhitungan (Matioli dkk., 2017).

Persamaan umum untuk menghitung bobot KDE:

$$KDE_i = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{\hat{x} - \hat{x}_i}{h}\right) \quad (2.6)$$

Keterangan:

$n$  : banyak data atau fitur

$\hat{x}$  : nilai probabilitas titik data dihitungnya KDE

$\hat{x}_i$  : nilai probabilitas data ke- $i$  dari  $n$  titik yang digunakan untuk estimasi  
 Penduga dari  $\hat{x}$  dihitung dengan menggunakan persamaan berikut:

$$\hat{x} = \frac{m_i}{M} \quad (2.7)$$

Keterangan:

$m_i$  : jumlah kemunculan kata  $i$  dalam korpus

$M$  : jumlah kemunculan semua kata dalam korpus

Kemudian penduga dari  $\hat{x}_i$  dihitung dengan menggunakan persamaan berikut:

$$\hat{x}_i = \frac{m_i}{N} \quad (2.8)$$

Keterangan:

$N$  : banyaknya dokumen

Untuk  $h$  adalah *bandwidth* menggunakan fungsi *Maximum Likelihood Estimation* (Silverman, 2018a).

$$h = 0,01 \sigma n^{-\frac{1}{5}} \quad (2.9)$$

Keterangan:

$\sigma$  : sigma adalah standar deviasi dari data

untuk  $K$  adalah kernel *Gaussian function*.

$$K = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\hat{x}-\hat{x}_i)^2}{2h^2}\right) \quad (2.10)$$

Sehingga, bobot KDE untuk setiap fitur akan menjadi:

$$KDE_i = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^n \exp\left(-\frac{(\hat{x} - \hat{x}_i)^2}{2h^2}\right) \quad (2.11)$$

Bobot KDE perlu dilakukan normalisasi menjadi nilai probabilitas termasuk menerapkan *smoothing* untuk menghindari bobot nol pada fitur data uji yang tidak terdapat pada data latih kemudian dapat digunakan sebagai distribusi probabilitas dalam fungsi *likelihood* (Silverman, 2018b). Rumus umum untuk melakukan *add-k smoothing* pada normalisasi bobot KDE adalah sebagai berikut:

$$P(KDE_i) = \frac{KDE_i + k}{\sum_i^n KDE_i + kn} \quad (2.12)$$

dengan  $k$  adalah nilai konstanta yang biasanya diganti dengan 1, dan  $n$  adalah jumlah data atau fitur.

## 2.8 Uji Persentase performa

Uji persentase performa adalah metode untuk mengevaluasi kinerja dari model *text mining*, di antaranya adalah:

1. Akurasi untuk mengukur seberapa baik model dapat mengklasifikasikan teks sesuai dengan label yang sebenarnya.
2. *Precision* untuk kemampuan model dapat mengklasifikasikan teks sebagai positif sesuai dengan label yang sebenarnya.
3. *Recall* untuk kemampuan menemukan semua teks positif yang sebenarnya.
4. *F1-Score* untuk mengukur keseimbangan antara *precision* dan *recall*.
5. *Confusion Matrix* untuk menghitung jumlah prediksi yang benar dan salah.

*Confusion matrix* adalah tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah dan memberikan informasi perbandingan hasil klasifikasi yang dilakukan dengan klasifikasi sebenarnya (Wibowo dkk., 2021).

*Confusion matrix* ditunjukkan pada **Tabel 2.1**:

**Tabel 2.1** *Confusion Matrix*

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	<i>TP</i>	<i>FN</i>
Negatif	<i>FP</i>	<i>TN</i>

Berikut merupakan penjelasan dari masing-masing nilai pada *confusion matrix*:

1. *TP (True Positive)* adalah jumlah data positif yang terklasifikasi dengan benar sebagai positif.
2. *FN (False Negative)* adalah jumlah data positif yang terklasifikasi salah sebagai negatif.
3. *FP (False Positive)* adalah jumlah data negatif yang terklasifikasi salah sebagai positif.
4. *TN (True Negative)* adalah jumlah data negatif yang terklasifikasi dengan benar sebagai negatif.

Berdasarkan pada nilai-nilai yang merupakan representasi kinerja klasifikasi tersebut dapat dihitung:

1. *Accuracy*

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (2.13)$$

2. *Recall*

$$Recall = \frac{TP}{TP + FN} \quad (2.14)$$

3. *Precision*

$$Precision = \frac{TP}{TP + FP} \quad (2.15)$$

4. *F<sub>measure</sub>*

$$F_{measure} = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (2.16)$$