ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

Klasifikasi Dokumen Tugas Akhir Berbasis *Text Mining* menggunakan Metode *Naïve Bayes Classifier* dan *K-Nearest Neighbor*

[1]Hamdani Asril, [2]Mustakim, [3]Insanul Kamila,
[1,2,3]Puzzle Research Data Technology (Predatech), Fakultas Sains dan Teknologi
Universitas Islam Negeri Sultan Syarif Kasim Riau
[1,2,3]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi
Universitas Islam Negeri Sultan Syarif Kasim Riau
Jl. HR. Soebrantas Km. 18 Panam Pekanbaru – Riau
e-mail: [1]hamdanisixx@gmail.com, [2]mustakim@uin-suska.ac.id
[3]insanulkamila17@gmail.com

Abstrak

Tugas akhir merupakan salah satu syarat yang harus dipenuhi mahasiswa untuk menyelesaikan pendidikan di perguruan tinggi. Dalam proses pengerjaan tugas akhir, mahasiswa membutuhkan dosen pembimbing sebagai tempat berkonsultasi sesuai dengan kategori/topik tugas akhir yang diajukan mahasiswa dalam proposal tugas akhir. Pada Program Studi Sistem Informasi UIN SUSKA Riau, penentuan dosen pembimbing tugas akhir ditentukan oleh Ketua Program Studi dengan melakukan kajian terstruktur berdasarkan judul/topik tugas akhir yang diajukan mahasiswa dengan keahlian bidang dosen yang sesuai. Untuk itu, perlu adanya sistem yang dapat memberikan rekomendasi agar pembagian dosen pembimbing lebih efisien. Penelitian ini bertujuan untuk membangun sistem rekomendasi dosen pembimbing menggunakan metode klasifikasi, Naïve Bayes Clasifier dan K-Nearest Neighbor (KNN). Dari percobaan 3 kelas dan 16 kelas diperoleh akurasi terbaik pada percobaan 3 kelas dengan nilai 86,11% untuk Naive Bayes Classifier (NBC) dan 91,67% untuk K-Nearest Neighbor (KNN). Pembangunan sistem menggunakan metode KNN untuk proses klasifikasinya dengan bahasa pemograman Python.

Kata kunci: *Klasifikasi, K-Nearest Neighbor, Naïve Bayes Clasifier, Text Mining, Tugas Akhir.*

Abstract

Thesis is one of the requirements that have to be applied by students to complete their education in a college. In the process of making a thesis, students need an academic advisor according to the categories of the thesis topic that have been submitted by students as a final project proposal. In the Information Systems Department of UIN SUSKA Riau, the academic advisors will determine based on the head of department by conducting a structured study based on the topic of the thesis submitted by students with expertised academic advisor. Therefore, it needs to build a system that can provide recommendations for academic advisors to be more efficient. The purpose of this study was to develop an academic advisor recommendation system by using the classification method called Naïve Bayes Classifier and K-Nearest Neighbor (KNN). From the experiments of 3 and 16 classes obtained that the best accuracy in the 3 classes experiment with a value of 86.11% by using Naive Bayes Classifier (NBC) and 91.67% for K-Nearest Neighbor (KNN). KNN method was used to develop the program and Python method was used for the programming language.

Keywords: *Classification, K-Nearest Neighbor, Naïve Bayes Clasifier, Text Mining, Undergraduate Thesis.*

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

1. Pendahuluan

Tugas akhir merupakan salah satu syarat yang harus dipenuhi oleh mahasiswa untuk menyelesaikan pendidikan di perguruan tinggi. Ilmu yang diperoleh mahasiswa dituangkan ke dalam suatu penelitian yang nantinya akan menghasilkan keluaran berupa dokumen tugas akhir[1]. Dalam proses pengerjaan tugas akhir, mahasiswa membutuhkan dosen pembimbing sebagai media berkonsultasi[2]. Dosen pembimbing tugas akhir memiliki peran penting karena memiliki tanggung jawab untuk memastikan bahwa mahasiswa mampu menyusun tugas akhir dengan baik hingga dapat menghasilkan tugas akhir yang berkualitas dan siap untuk diuji[3].

Untuk itu dibutuhkan dosen pembimbing yang tepat dengan bidang keahlihan sesuai dengan kategori/ bidang ilmu topik tugas akhir yang diajukan mahasiswa dalam proposal tugas akhir. Hal ini bertujuan agar konsep dan perancangan tugas akhir yang dibuat dalam proposal dapat dikerjakan dan diwujudkan sesuai dengan tujuan yang direncanakan, dengan harapan tidak adanya kendala[4]. Pada Program Studi Sistem Informasi Universitas Islam Sultan Syarif Kasim Riau (UIN SUSKA Riau) penentuan Dosen pembimbing ditentukan berdasarkan keputusan dari Ketua Program Studi (Kaprodi). Kaprodi menentukan dosen pembimbing berdasarkan topik tugas akhir mahasiswa dan keahlian dosen. Penentuan dosen pembimbing tugas akhir ditentukan oleh Kaprodi secara manual berdasarkan keahlian bidang dosen pada Sistem Informasi sehingga Ketua Program Studi melakukan kajian yang terstruktur antara judul/topik tugas akhir yang diajukan mahasiswa dengan keahlian bidang dosen yang sesuai. Untuk itu, perlu adanya terobosan baru untuk meminimalisir waktu agar lebih efisien dan efektif dalam proses penentuan dosen pembimbing.

Penelitian ini bertujuan untuk mempermudah Kaprodi dalam menentukan Dosen pembimbing bagi Mahasiswa yang mengajukan Proposal Tugas Akhir. Untuk melihat sesuai atau tidaknya judul Tugas Akhir dengan Dosen pembimbing maka dilakukan *text mining* menggunakan algoritma *Naïve Bayes Clasifier* (NBC) dan *K-Nearest Neighbor* (KNN) dengan melihat kesesuaian judul-judul tugas akhir sebelumnya yang pernah dibimbing oleh pembimbing tersebut. *Text mining* memiliki peran penting dalam bidang data mining. Dengan mengaplikasikan proses-proses dalam *text mining*, maka akan diperoleh pola-pola data, tren, dan ekstraksi dari pengetahuan-pengetahuan yang potensial dari data teks[5].

Text mining merupakan salah satu bidang khusus dari *data mining*, yang artinya adanya interaksi antara *user* dengan dokumen-dokumen berkelanjutan dari waktu ke waktu dalam proses penggalian informasi menggunakan seperangkat *tools* analisis. *Text mining* merupakan teknik yang dapat digunakan untuk menyelesaikan masalah klasifikasi, *clustering, information extraction* dan *information retrieval* untuk mengekstrak informasi yang diambil dari sumber-sumber yang berbeda. Dikarenakan kebanyakan informasi (perkiraan umum mengatakan bahwa mencapai lebih dari 80%) saat ini disimpan sebagai teks, maka *text mining* diyakini memiliki potensi nilai komersial tinggi[6].

Naïve Bayes Classifier (NBC) merupakan salah satu algoritma klasifikasi untuk mengklasifikasikan sekumpulan teks kedalam kelas yang ada. Algoritma ini memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya [7]. Selain menggunakan NBC, pada penelitian ini juga menggunakan algoritma *K-Nearest Neighbor* (KNN) yang merupakan metode klasifikasi untuk mengklasifikasikan objek kedalam kelas tertentu berdasarkan data latih yang jaraknya paling dekat dari objek tersebut [8]. Prinsip kerja dari KNN adalah mencari jarak antara dua titik yaitu titik *training* dan titik *testing*. Selanjutnya dilakukan evaluasi berdasarkan k tetangga terdekatnya dalam data *training*[1].

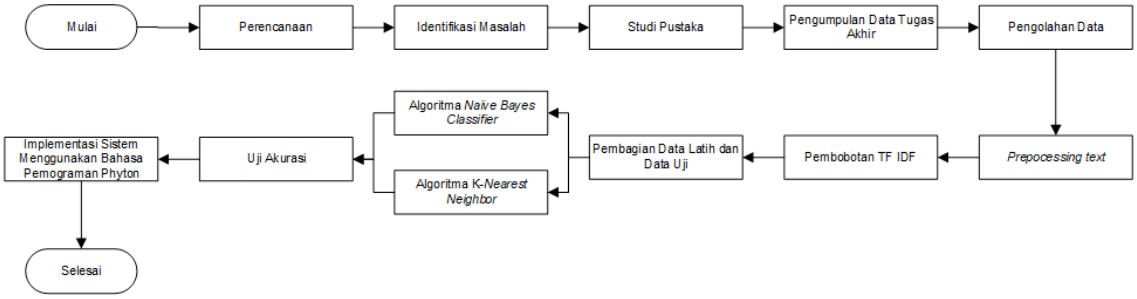
NBC memberikan hasil yang lebih tinggi dibandingkan dengan SVM pada penelitian Hidayatullah dan Ma'arif untuk Penerapan *Text Mining* dalam Klasifikasi Judul Skripsi. Dilihat dari penelitian sebelumnya bahwasannya NBC mampu memberikan hasil yang lebih tinggi dibandingkan dengan metode lainnya, untuk itu metode NBC akan diimplementasilam pada penelitian ini untuk mendapatkan hasil yang maksimal[5]. Selanjutnya pada penelitian Sani, Zeniarja, dan Luthfiarta(2016) menghasilkan akurasi sebesar 80% menggunakan algoritma KNN dalam Penentuan Topik Referensi Tugas Akhir berupa data abstrak tugas akhir mahasiswa yang didapat dari berbagai sumber[22].

Dilihat dari penelitian sebelumnya, bahwasannya NBC dan KNN mampu menghasilkan klasifikasi yang sesuai. Kelebihan penelitian ini dibandingkan penelitian sebelumnya yaitu pada penelitian ini akan membandingkan algoritma NBC dan KNN. Algoritma terbaik akan diterapkan kedalam sistem yang dibangun yaitu sistem klasifikasi dosen pembimbing tugas akhir menggunakan bahasa pemograman phyton.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

2. Metodologi

Adapun metodologi yang diterapkan dalam melakukan penelitian ditunjukkan pada gambar berikut:



Gambar 1. Metodologi Penelitian

2.1 Text Mining

Text mining merupakan salah satu bidang khusus dari data mining. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* ruang lingkup yang besar[10]. *Text mining* adalah satu langkah dari analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas terangkum dalam sebuah dokumen dari berbagai rangkaian teks[11]. *Text mining* merupakan teknik yang digunakan untuk menyelesaikan masalah klasifikasi, *clustering*, *information extraction* dan *information retrieval*[13]. Prosedur utama dalam metode ini terkait dengan menemukan kata-kata yang dapat mewakili isi dari dokumen untuk selanjutnya dilakukan analisis keterhubungan antar dokumen dengan menggunakan metode statistik tertentu seperti analisis kelompok, klasifikasi dan asosiasi. Adapun tahapan awal dalam *text mining* disebut dengan *text preprocessing*.

2.2 Klasifikasi

Klasifikasi merupakan kegiatan dalam mengukur objek data untuk dikelompokkan ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi, terdapat dua jenis pekerjaan utama yang dilakukan, yaitu (1) pembangunan model sebagai *prototype* untuk disimpan sebagai memori dan (2) penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data baru agar dapat diketahui kelas mana objek data tersebut berdasarkan model yang sudah disimpan[14].

2.3 *Naïve Bayes Classifier (NBC)*

Algoritma NBC merupakan metode klasifikasi statistik yang didasarkan oleh teorema Bayes[15]. NBC merupakan pengklasifikasian sebuah metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu dapat memprediksi kemungkinan di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan *Naive* dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi *Naive Bayes* diasumsikan bahwa ada atau tidaknya suatu ciri tertentu dari sebuah kelas dan tidak memiliki hubungan dengan ciri dari kelas lainnya. NBC berpotensi cukup tinggi untuk mengklasifikasikan data karena kesederhanaannya[16].

Persamaan NBC untuk klasifikasi adalah[2] :

$$p(W_i|C_j) = \frac{N_{cw}+1}{N_c+V}$$

Dimana :

Ncw = jumlah kata wi yang ada dalam dokumen *training* yang masuk ke dalam kategori Cj.

Nc = jumlah semua kata yang ada dalam dokumen *training* yang masuk kedalam kategori Cj (tanpa menghiraukan ada kata yang sama atau tidak).

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

V = jumlah total jenis kata yang ada dalam dokumen training (kata yang sama hanya dihitung 1).

Prinsip dari NBC adalah probabilitas suatu kata akan masuk ke dalam suatu kategori (*posterior probability*), didasarkan pada nilai probabilitas tertinggi yang telah dimiliki sebelumnya (*prior probability*), yang dimiliki teks yang bersangkutan untuk suatu kategori tertentu. Misalnya kata "*processor*" pada koleksi data memiliki probabilitas untuk kategori "komputer" sebesar 0.9, sedangkan untuk kategori "elektronik" sebesar 0.3. Sehingga pada proses pengujian ditemukan kata "processor" maka akan masuk ke dalam kategori "komputer". Dengan kata lain, NBC menggunakan asumsi bahwa kemunculan atau ketidakmunculan dari suatu kata atau fitur tidak terkait dengan kemunculan atau ketidakmunculan fitur yang lain [11].

2.4 *K-Nearest Neighbor (KNN)*

Algoritma ini pertama kali diperkenalkan oleh Fix dan Hodges pada tahun 1951 dan 1952[17]. Algoritma ini juga merupakan salah satu teknik *lazy learning*, dimana KNN dilakukan dengan mencari kelompok k objek dalam data *training* yang paling dekat (mirip) dengan objek pada data baru atau data *testing* [18][21].

KNN mengklasifikasikan objek berdasarkan data latih yang jaraknya paling dekat dari objek tersebut. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak *euclidean* dengan persamaan umum [9]:

$$d = \sqrt{\sum_{i=1}^{n} (a_i - b_1)^2}$$

Keterangan:
d = jarak
a = data uji/*testing*
b = sampel data
i = variable data
n = Dimensi data

2.5 *Confusion Matrix*

Confusion matrix merupakan sebuah model evaluasi klasifikasi berdasarkan data uji dan seluruh data yang diprediksi dengan proporsi yang tepat [19].

Tabel 1. Tabel *Confusion Matrix* 2 Kelas

Classification	Prediction Class	
	Yes	No
Yes	A (True Positive)/TP	B (False Postive)/FN
No	C (False Postive)/FP	D (True Negative)/TN

Adapun perhitungan tingkat akurasi pada Confusion Matriks 2 kelas berdasarkan persamaan 3 [20] adalah :

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{A + D}{A + B + C + D}$$

Pada penelitian ini dilakukan beberapa simulasi nilai k pada metode KNN dan melakukan perbandingan metode NBC dan KNN untuk mendapatkan metode dengan akurasi terbaik di penelitian ini. Akurasi dihitung dengan *Confusion Matrix.*

3. Analisa dan Hasil

3.1 Analisa Kebutuhan Data

Data yang digunakan pada penelitian ini yaitu data Tugas Akhir (TA) program studi Sistem Informasi Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim (UIN SUSKA) Riau tahun 2013 sampai tahun 2018. Total data keseluruhan sebanyak 480 judul TA.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

3.2 Penentuan Atribut

Atribut atau kriteria yang dipilih sesuai dengan kebutuhan dan yang berhubungan dengan penelitian. Adapun atribut yang digunakan pada penelitian ini yaitu judul TA dan nama pembimbing dari data seminar TA.

3.3 *Cleaning*

Setelah data diperoleh dan atribut ditentukan, selanjutnya melakukan pembersihan atau *cleaning* data. pembersihan data dilakukan utuk mengurangi kerancuan dan *noise* yang dapat mempengaruhi perhitungan. Pembersihan data dilakukan dengan cara menghapus data yang tidak memiliki nilai pada setiap atribut serta *duplicate* data, terdapat 58 *record* data yang kosong di setiap atribut dan *duplicate* data. Selanjutnya dilakukan penghapusan untuk 58 *record* data tesebut sehingga jumlah data setelah dilakukan *cleaning* sebanyak 422 *record* data untuk data seminar TA.

3.4 *Text Preprocessing*

Tahap *preprocessing* atau praproses data merupakan proses untuk mempersiapkan data mentah sebelum dilakukan analisis data. Pada umumnya, praproses data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. Praproses sangat penting dalam melakukan analisis data. Pada tahapan ini *preprocessing* menggunakan *tools jupyter notebook* dengan bahasa pemrograman *python* untuk memperoleh data siap proses.

Tahap pertama yaitu *tokenizing* yaitu proses penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan delimiter-delimiter seperti tanda titik (.), koma (,), spasi dan karakter angka yang ada pada kata tersebut (Weiss dkk, 2005). Setelah melakukan *tokenizing* selanjutnya yaitu proses *filtering* dimana seleksi terhadap kata-kata yang dihasilkan dari proses *tokenizing* dengan algoritma *stopword*. Algoritma *stopword* akan membuang kata-kata yang tidak penting seperti kata ganti, kata keterangan, kata sambung, kata depan dan kata sandang. Daftar kata *stopword* di penelitian ini bersumber dari Tala (2003). Berikutnya merupakan tahapan *stemming* salah satu algoritma yang mendukung *stemming* bahasa Indonesia yaitu *confix-stripping* dimana kata-kata berimbuhan diubah menjadi bentuk dasar.

3.5 Data *Training* dan *Testing*

Data latih (*data training*) dan data uji (*data testing*) diambil dari data awal dengan jumlah 422 judul TA setelah melalui beberapa tahapan. Data tersebut dibagi untuk dilakukan pengujian menggunakan algoritma *Naïve Bayes Classifier* (NBC) dan *k-Nearest Neighbor* (KNN). Pada data judul TA akan dilakukan percobaan perhitungan menggunakan data 3 kelas dan seluruh kelas. Untuk penggunaan data 3 kelas dipilih karena tiap kelas dari dosen yang terpilih memiliki data terbanyak dan seimbang dari 422 data awal, sehingga pembobotan kelas yang dilakukan memiliki *range* nilai yang tidak jauh berbeda dengan masing-masing kelas dosen. Selanjutnya menggunakan seluruh kelas dengan jumlah dosen keseluruhan yaitu sebanyak 16 kelas. Data judul TA dibagi menjadi data *training* dan data *testing* dengan perbandingan 70:30, menggunakan modul sklearn yaitu train_test_split pada python.

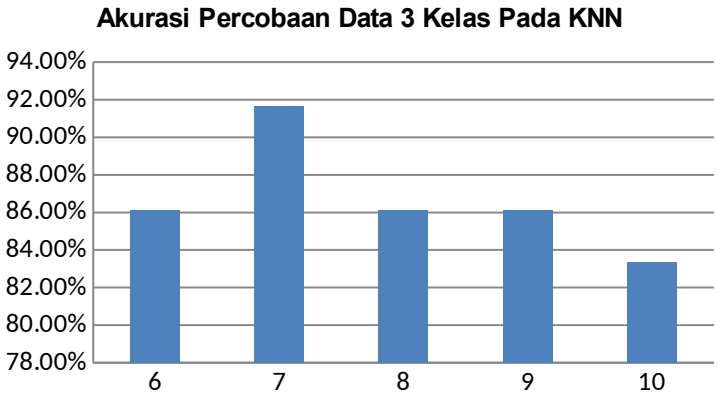
3.6 *Term Frequency-Inverse Document Frequency* (TF-IDF)

Perhtiungan TF-IDF pada penelitian ini menggunakan bahasa pemrograman python dengan menggunakan modul scikit-learn yaitu TfidfVectorizer. TF-IDF dibagi sesuai dengan pembagian data *training* dan data *testing.* TF-IDF dihitung berdasarkan data percobaan yaitu TF-IDF pada data menggunakan 3 kelas dan TF-IDF pada data menggunakan 16 kelas.

3.7 Hasil *Naïve Bayes Classifier* (NBC) dan *K-Nearest Neighbor*

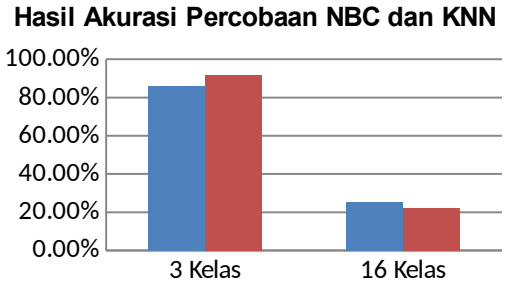
Berikut adalah hasil percobaan klasifikasi NBC menggunakan module scikit-learn pada python yaitu MultiNomialNB dan klasifikasi KNN menggunakan module scikit-learn pada python yaitu KNeighborsClassifier. Pada perhitungan KNN akan dilakukan percobaan nilai k menggunakan nilai k=6, 7, 8, 9 dan 10 dan didapatkan hasil k=7 terbaik, perbadingan nilai k dapat dilihat seperti pada gambar 2 berikut:

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406



Gambar 2. Akurasi Percobaan Nilai Pada Data 3 Kelas

Hasil akurasi kedua algoritma dapat dilihat pada Gambar 3 berikut:



Gambar 3. Akurasi Hasil Percobaan NBC dan KNN

3.8 Analisis Klasifikasi

Data uji dengan menggunakan 3 kelas memiliki jumlah data secara keseluruhan 120 dengan jumlah data disetiap kelasnya yaitu Syaifullah, Se, M.Sc 40 data, Eki Saputra, S.Kom, M.Kom 42 data dan Mustakim, ST, M.Kom 38 data. Akurasi yang diperoleh pada perhitungan NBC yaitu 86,11% sedangkan pada KNN akurasi yang diperoleh lebih besar yaitu 91,67% dengan nilai k=7. Dengan demikian algoritma KNN memiliki akurasi lebih besar pada percobaan menggunakan data 3 kelas ini.

Selanjutnya pada data uji dengan menggunakan 16 kelas atau kelas secara keseluruhan yang ada pada dataset memiliki jumlah data sebanyak 422 data dengan jumlah data disetiap kelasnya bervariasi mulai dari yang terbesar yaitu 42 data hingga yang terkecil yaitu 7 data. Akurasi yang diperoleh pada perhitungan NBC yaitu 25,20% lebih besar dibandingkan akurasi yang diperoleh dari perhitungan KNN yaitu 22,05% dengan nilai k=8. Dengan demikian algoritma NBC memiliki akurasi lebih besar pada percobaan menggunakan data 16 kelas atau kelas secara keseluruhan ini.

Dari analisis diatas dapat diambil kesimpulan bahwasannya akurasi yang baik dapat diperoleh jika sebaran data disetiap kelas tidak memiliki range yang terlalu jauh seperti pada percobaan menggunakan 3 kelas.

3.9 Batasan Implementasi Sistem Aplikasi

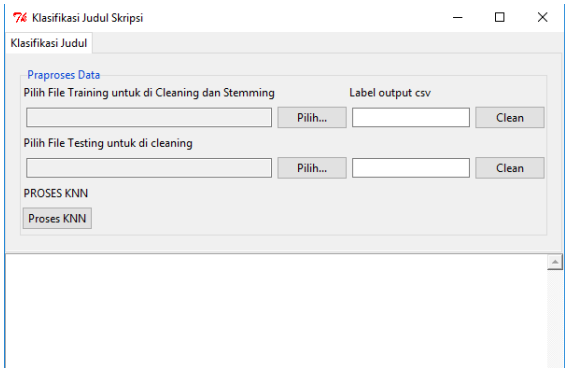
Batasan implementasi pada penelitan ini adalah sebagai berikut:

1. Aplikasi yang dibangun berbasis desktop.
2. Aplikasi bersifat open atau publik.
3. Aplikasi dibangun menggunakan bahasa pemrograman python dengan library GUI Tkinter.
4. Aplikasi tidak menggunakan database, data input dan output adalah file dengan format csv.
5. Aplikasi dapat melakukan cleaning dan stemming data, serta melakukan hitungan klasifikasi KNN.
6. Output akhir dari aplikasi adalah nilai probabilitas dosen pembimbing Tugas Akhir.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

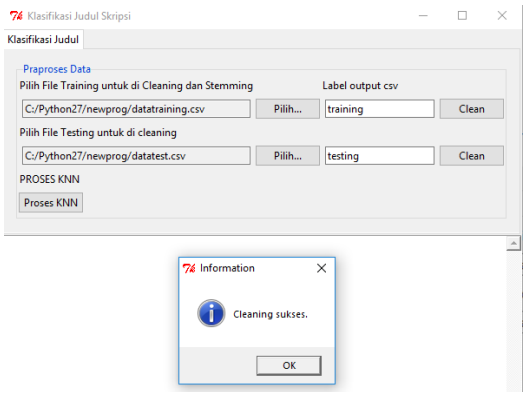
3.10 Implementasi Sistem Aplikasi

Implementasi dari aplikasi dapat dilihat pada Gambar 4 berikut:



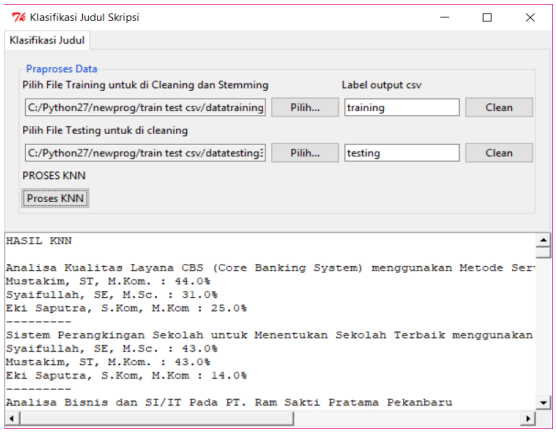
Gambar 4. Implementasi Aplikasi

Gambar 3 merupakan tampilan implementasi aplikasi berbasis desktop yang dibuat menggunakan bahasa pemrograman Python. Untuk menjalankan aplikasi langkah pertama yaitu memilih data *training* dan data *testing* yang akan dihitung. Selanjutnya mengisi *label output* atau nama *file* dari hasil *cleaning* nntinya. Setelah itu tekan *button* "*Clean*", maka akan keluar tampilan seperti pada Gambar 4 berikut:



Gambar 4. Tampilan *Preprocessing*

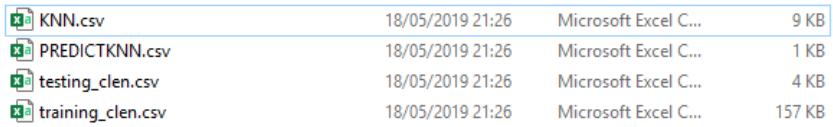
Setelah melakukan *cleaning*, selanjutnya ke proses perhitungan KNN. Hasil perhitungan akan diperoleh saat *button* "Proses KNN" ditekan. Maka muncul hasil perhitungan KNN seperti pada Gambar 5 berikut ini:



Gambar 5. Tampilan Hasil KNN

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

File hasil dari *cleaning* dan perhitungan KNN dapat dilihat pada penyimpanan sesuai direktori yang ditentukan, seperti pada Gambar 6 berikut ini:



Gambar 6. Tampilan *File* Hasil

4. Kesimpulan

Adapun setelah didapatkan hasil penelitian ini, kesimpulan yang didapatkan yaitu:

1. Dari percobaan 3 kelas dan 16 kelas diperoleh akurasi terbaik pada percobaan 3 kelas dengan nilai 86,11% untuk *Naive Bayes Classifier* (NBC) dan 91,67% untuk *K-Nearest Neighbor* (KNN) dan percobaan 16 kelas dengan nilai 25,20% untuk *Naive Bayes Classifier* (NBC) dan 22.05% untuk *K-Nearest Neighbor* (KNN).
2. Pada jumlah kelas yang sedikit menggunakan percobaan 3 kelas dengan jumlah data yang sedikit yaitu 120 data, algoritma KNN menghasilkan akurasi yang lebih tinggi dibandingkan algoritma NBC. Namun pada jumlah kelas yang banyak menggunakan percobaan seluruh kelas yaitu 16 kelas dengan jumlah data lebih banyak yaitu 422 data, NBC menghasilkan akurasi yang cukup tinggi dibandingkan dengan KNN. Sehingga, dalam implementasi sistem menggunakan bahasa pemograman Pyhton dipilih algoritma KNN pada percobaan 3 kelas karena memiliki akurasi yang lebih tinggi.
3. Pada algoritma KNN nilai k yang memiliki akurasi tertinggi yaitu k=6, k=7, k=8 dan k=9, dengan nilai k=7 sebesar 91,67% dan sebesar 86,11% untuk nilai akurasi yang sama pada nilai k lainnya.
4. Akurasi yang bagus dapat diperoleh jika sebaran data disetiap kelas tidak memiliki range yang terlalu jauh seperti pada percobaan menggunakan 3 kelas.

Daftar Pustaka

[1] Yusra., Olivita, D., dan Vitriani, Y. “Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode Naïve Bayes Classifier dan K-Nearest Neighbor”. Jurnal Sains, Teknologi dan Industri, vol. 14 no. 1. 2016.

[2] Mas’udia, P. E. “Klasifikasi Tugas Akhir Untuk Menentukan Dosen Pembimbing Menggunakan Naïve Bayes Classifier (NBC)”. Prosiding SENTIA, Vol. 7. 2015.

[3] Hariyati, R. M. “Survey Kinerja Dosen Pembimbing Skripsi dan Kualitas Skripsi Mahasiswa Akhuntansi STIE Malangkucecwara”. Jurnal Dinamika Akuntansi, Vol. 4, No. 2. 2012.

[4] Kasih, P. “Integrasi Kategori Skripsi dan Keahlian Dosen Dalam Naive Bayes Untuk Pemilihan Dosen Pembimbing”. Nusantara of Engineering, Vol. 3, No.2. 2016.

[5] Hidayatullah, A.F., dan Ma’arif, M.R. “Penerapan Text Mining dalam Klasifikasi Judul Skripsi”. Seminar Nasional Aplikasi Teknologi Informasi (SNATi). 2016.

[6] Bridge, C. Unstructured Data and the 80 Percent Rule. [Online] Available. http://www.clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=55.Diakses 24 Mei 2018.

[7] Noah, S. A., dan Ismail, F. *“Automatic Classifications of Malay Proverbs Using Naive Bayesian Algorithm”. Information Technology Journal.* 2008.

[8] Larasati, R. Klasifikasi Teks Dengan Menggunakan Algoritma K-Nearest Neighbor Pada Dokumen Tugas Akhir. [Skripsi] Universitas Widyatama. 2015.

[9] Efendi Z., dan Mustakim. “Text Mining Classification Sebagai Rekomendasi Dosen Pembimbing Tugas Akhir Program Studi Sistem Informasi”. Seminar Nasional Teknologi Informasi, Komunikasi dan Industri (SNTIKI) 9. 2017.

[10] Turban, E., J. E. Aronson, dan T. P. Liang. *Introduction to Data Mining. Pearson.* 2005.

[11] Han, J., & Kamber, M. *Data Mining Concepts and Techniques Second Edition.* San Fransisco: Morgan Kaufmann Publisher. 2006.

[12] Feldman, R., dan J. Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.* New York. 2007.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

[13] Berry, M. W., dan J. Kogan. *Text Mining Application and Theory.* WILEY. United Kingdom. 2010.

[14] Prasetyo, E. Data Mining Konsep dan Aplikasi menggunakan Matlab. Penerbit Andi. Yogyakarta. 2012.

[15] Baby, N. *"Customer Classification And Prediction Based On Data Mining Technique". International Journal of Emerging Technology and Advanced Engineering*. Vol. 2. 2012.

[16] Ting, S.L., Ip, W.H., dan Tsang, A.H.C. *"Is Naïve Bayes a Good Classifier for Document Classification?,". International Journal of Software Engineering and Its Applications*. Vol. 5. 2011.

[17] Santoso dan Irawan, M. I. "*Classification of Poverty Levels using KNearest Neighbor". InternationI Journal of Computing and Science and Applied Mathematics,* vol. 2, No 1. 2012.

[18] Leiydiana, H. "Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor". Jurnal Penelitian Ilmu Komputer, System Embedded & Logic, Vol: 01, 65-76. 2013.

[19] Faiza, N. N. "Prediksi Tingkat Keberhasilan Mahasiswa Tingkat I IPB dengan Metode k-Nearest Neighbor". Institut Pertanian Bogor, Bogor, Indonesia: Institut Pertanian Bogor. 2009.

[20] Adriani, M., J. Asian, B. Nazief, S. M.M. Tahaghoghi, dan H. E. Williams. "*Stemming Indonesian: A Confix-Stripping Approach*." Transaction on Asian Langeage Information Processing Vol. 6 No. 4. 2007.

[21] Okfalisa, Gazalba, I., Mustakim, Reza, N.G.I. Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification. Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE. 2018.

[22] Sani, RR., Zeniarja, J., Luthfiarta, A. Penerapan Algoritma K-*Nearest Neighbor* pada *Information Retrieval* dalam Penetuan Topik Referensi Tugas Akhir. *Journal of Applied Intelligent System,* Vol. 1, No.2, Juni 2016; 123-133.