KLASIFIKASI BIDANG ILMU PADA PUBLIKASI TERINDEKS *WEB OF SCIENCE* MENGGUNAKAN METODE *K-NEAREST NEIGHBOR*

LAPORAN TUGAS AKHIR

Laporan Ini Disusun Guna Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Strata (S1) pada Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang

Disusun Oleh :

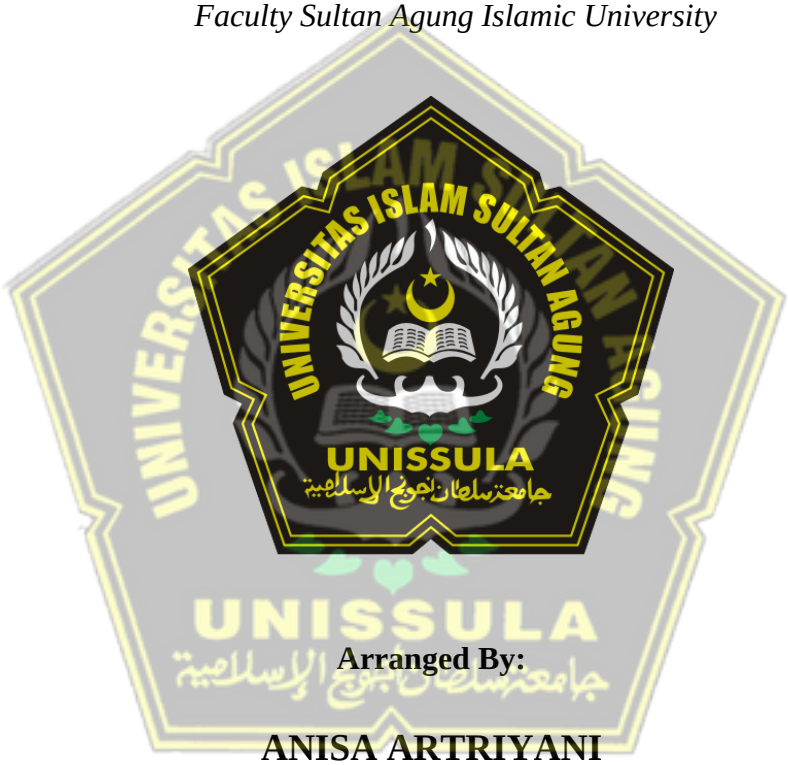
ANISA ARTRIYANI
NIM 32601800008

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ISLAM SULTAN AGUNG
SEMARANG
DESEMBER 2022

FINAL PROJECT

CLASSIFICATION OF SCIENCE FIELDS IN WEB OF SCIENCE INDEXED PUBLICATIONS USING THE K-NEAREST NEIGHBOR METHOD

Proposed to complete the requirement to obtain a bachelor's degree (S-1) at Informatics Engineering Departement of Industrial Technology Faculty Sultan Agung Islamic University



Arranged By:

ANISA ARTRIYANI
NIM 32601800008

MAJORING OF INFORMATICS ENGINEERING

INDUSTRIAL TECHNOLOGY FACULTY

SULTAN AGUNG ISLAMIC UNIVERSITY

SEMARANG

DECEMBER 2022

LEMBAR PENGESAHAN PEMBIMBING

Laporan Tugas Akhir dengan judul **"KLASIFIKASI BIDANG ILMU PADA PUBLIKASI TERINDEKS *WEB OF SCIENCE* MENGGUNAKAN METODE *K-NEAREST NEIGHBOR*"** ini disusun oleh :

Nama :Anisa Artriyani
NIM :32601800008
Program Studi : Teknik Informatika
Telah disahkan oleh dosen pembimbing pada :
Hari :
Tanggal :

Mengesahkan,

Pembimbing I	Pembimbing II
Sam Farisa C.H, ST,M.Kom	Andi Riansyah,ST,M.Kom
NIDN. 0628028602	NIDN.0609108802

Mengetahui,

Ketua Program Studi Teknik Informatika
Fakultas Teknologi Industri
Universitas Islam Sultan Agung

Ir.Sri Mulyono,M.Eng
NIDN.0623117703

LEMBAR PENGESAHAN PENGUJI

Laporan tugas akhir dengan judul **"Klasifikasi Bidang Ilmu pada Publikasi Terindeks *Web of Science* Menggunakan Metode *K-Nearest Neighbor*"** ini telah dipertahankan di depan dosen penguji Tugas Akhir pada :

Hari :

Tanggal :

TIM PENGUJI

Penguji I	Penguji II
Badie'ah, ST,M.Kom	Ghufron, ST,M.Kom
NIDN.0619018701	NIDN. 0602079005

SURAT PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan dibawah ini :

Nama : Anisa Artriyani

NIM : 32601800008

Judul Tugas Akhir : Klasifikasi Bidang Ilmu pada Publikasi Terindeks *Web of Science* menggunakan Metode *K-Nearest Neighbor*.

Dengan bahwa ini saya menyatakan bahwa judul dan isi Tugas Akhir yang saya buat dalam rangka menyelesaikan Pendidikan Strata Satu (S1) Teknik Informatika tersebut adalah asli dan belum pernah diangkat, ditulis ataupun dipublikasikan oleh siapapun baik keseluruhan maupun sebagian, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka, dan apbila di kemudian hari ternyata terbukti bahwa judul Tugas Akhir tersebut pernah diangkat, ditulis ataupun dipublikasikan, maka saya bersedia dikenakan sanksi akademis. Demikian surat pernyataan ini saya buat dengan sadar dan penuh tanggung jawab.

Semarang,25 Februari 2023

Yang menyatakan,

Anisa Artriyani

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Saya yang bertanda tangan dibawah ini :

Nama	: Anisa Artriyani
NIM	: 32601800008
Program Studi	: Teknik Informatika
Fakultas	: Teknologi Industri
Alamat Asal	: Bogorame RT 02 RW 01 Kel.Mangunjiwan Kec.Demak Kab. Demak.

Dengan ini menyatakan Karya Ilmiah berupa Tugas akhir dengan Judul : **Klasifikasi Bidang Ilmu pada Publikasi Terindeks *Web of Science* menggunakan Metode *K-Nearest Neighbor*.**

Menyetujui menjadi hak milik Universitas Islam Sultan Agung serta memberikan Hak bebas Royalti Non-Eksklusif untuk disimpan, dialihmediakan, dikelola dan pangkalan data dan dipublikasikan diinternet dan media lain untuk kepentingan akademis selama tetap menyantumkan nama penulis sebagai pemilik hak cipta. Pernyataan ini saya buat dengan sungguh-sungguh. Apabila dikemudian hari terbukti ada pelanggaran Hak Cipta/Plagiarisme dalam karya ilmiah ini, maka segala bentuk tuntutan hukum yang timbul akan saya tanggung secara pribadi tanpa melibatkan Universitas Islam Sultan agung.

Semarang, 25 Februari 2023

Yang menyatakan,

Anisa Artriyani

KATA PENGANTAR

Dengan mengucap syukur alhamdulillah atas kehadirat ALLAH SWT yang telah memberikan rahmat dan karunianya kepada penulis, sehingga dapat menyelesaikan Tugas Akhir dengan judul "Klasifikasi Bidang Ilmu pada Publikasi Terindeks Web Of Science Menggunakan Metode K-Nearest Neighbor" ini untuk memenuhi salah satu syarat menyelesaikan studi serta dalam rangka memperoleh gelar sarjana (S-1) pada Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang.

Tugas Akhir ini disusun dan dibuat dengan adanya bantuan dari berbagai pihak, materi maupun teknis, oleh karena itu saya selaku penulis mengucapkan terima kasih kepada:

1. Rektor UNISSULA Bapak Prof. Dr. H. Gunarto, SH, M.Hum. yang mengijinkan penulis menimba ilmu di kampus ini.
2. Dekan Fakultas Teknologi Industri Ibu Dr. Ir. Hj. Novi Marlyana, S.T, M.T.
3. Dosen pembimbing I penulis Sam Farisa C.H, ST, M.Kom yang telah meluangkan waktu dan memberikan ilmu.
4. Dosen pembimbing II penulis Andi Riansyah, ST, M.Kom yang memberikan banyak nasehat dan saran.
5. Orang tua penulis yang telah mengijinkan untuk menyelesaikan laporan ini serta dukungan materil.
6. Teman saya Sofianisa Fitriyati Prisunia dan Teman seperjuangan bimbingan yang telah membantu saya dalam menyelesaikan penulisan laporan tugas akhir, serta kepada semua pihak yang tidak dapat saya satu persatu.

Dengan segala kerendahan hati, penulis menyadari masih banyak terdapat kekurangan dalam penyusunan laporan, sehingga penulis mengharapkan adanya saran dan kritikan yang bersifat membangun demi kesempurnaan laporan ini dan masa mendatang.

Semarang, 25 Februari 2023

Anisa Artriyani

DAFTAR ISI


UNIVERSITAS ISLAM SULTAN AGUNG
UNISSULA
جامعة سلطان أجونج الإسلامية

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Web of Science (WoS) adalah *database* yang menawarkan pengindeksan kutipan dari publikasi bereputasi internasional. SINTA mengagregasi berbagai sumber publikasi baik internasional maupun nasional, salah satu publikasi internasional bereputasi adalah WoS. SINTA merupakan suatu sarana untuk mengkomunikasikan karya IPTEK manusia dalam sebuah bentuk berbasis web berisi sistem informasi publikasi penelitian yang di rintis Direktur Jenderal Penguatan Penelitian dan Pengembangan, Kementrian Riset Teknologi dan Dikti Republik Indonesia pada tahun 2016. Maka dari itu, penulis melakukan sebuah penelitian bagaimana mengklasifikasi data publikasi WoS sesuai dengan lima bidang ilmu. Penelitian ini bertujuan mengklasifikasi judul artikel sesuai dengan bidang ilmu terindeks dalam SINTA yang terdapat di WoS menggunakan metode *K-Nearest Neighbor*. Dengan teknik yang tepat, dapat memperoleh strategi dan prosedur yang akan dijalankan. Pada penelitian ini data diperoleh dengan lima bidang ilmu yakni *Art & Humanities, Engineering & Technology, Life Sciences & Medicine, Natural Sciences, Social Sciences & Management*. sampel data dari SINTA yang terindeks WoS yang berjumlah 1000, masing-masing data pada bidang ilmu yaitu 200. Dengan data Training 900 dan data Testing 100. Mendapatkan nilai akurasi sebesar 0.50, nilai *recall* yaitu sebesar 0.27, dan nilai presisi sebesar 0.21 dengan nilai K= 25, dimana hasil tersebut telah dilakukan beberapa kali percobaan. Hal tersebut menunjukan bahwa klasifikasi judul artikel publikasi terindeks WoS dengan metode KNN untuk penelitian ini masih belum sesuai harapan karena terdapat berbagai faktor penyebabnya antara lain data yang didapatkan kurang tepat dan akurat sehingga menghasilkan nilai akurasi yang diperoleh tidak sesuai yang diharapan.

Kata kunci : *Web Of Science*, SINTA, *K-Nearest Neighbor*..

ABSTRACT

Web of Science (WoS) is a database that offers indexing of citations from internationally reputable publications. SINTA aggregates various publication sources both international and national, one of the reputable international publications is WoS. SINTA is a means to communicate the work of human science and technology in a web-based form containing a research publication information system pioneered by the Director General of Strengthening Research and Development, Ministry of Research Technology and Higher Education of the Republic of Indonesia in 2016. Therefore, the author conducted a study on how to classify WoS publication data according to five fields of science. This research aims to classify article titles according to the field of science indexed in SINTA contained in WoS using the K-Nearest Neighbor method. With the right technique, it can obtain strategies and procedures that will be carried out. In this study, data was obtained with five fields of science namely Art & Humanities, Engineering & Technology, Life Sciences & Medicine, Natural Sciences, Social Sciences & Management. sample data from SINTA indexed by WoS which amounted to 1000, each data in the field of science is 200 with 900 Training data and 100 Testing data. Get an accuracy value of 0.50, a recall value of 0.27, and a precision value of 0.21 with a value of K = 25, where these results have been carried out several times. This shows that the classification of WoS indexed publication article titles with the KNN method for this research is still not as expected because there are various factors that cause it, among others, the data obtained is less precise and accurate, resulting in the accuracy value obtained is not as expected.

Keyword : *Web Of Science*, SINTA, *K-Nearest Neighbor*..

BAB I
PENDAHULUAN

1.1 Latar Belakang

Web of Science (WoS) merupakan database yang menawarkan pengindeksasian kutipan dari publikasi yang bereputasi internasional. Sebelumnya, Wos dikenal sebagai *Web of Knowledge* merupakan suatu database bibliografi pertama yang diciptakan Eugene Garfield pada tahun 1960 sebagai *Institute for Scientific Information* (ISI). Tahun 2016, Wos diakuisisi oleh *Clarivate Analytics*. Wos sebagai database selektif yang terdiri dari berbagai indeks khusus, dikelompokkan berdasarkan jenis tema terindeks.

Science and Technology Index atau SINTA adalah media yang memungkinkan untuk mendukung Ilmu dan Pengetahuan Teknologi (IPTEK) manusia dalam sistem pengumpulan informasi berbasis *online* yang dilaksanakan Kementrian Riset Teknologi dan Dikti Republik Indonesia pada tahun 2016. Web SINTA memiliki sebuah visi untuk menjadi tolak ukur bagi pengkajian di Indonesia.

SINTA sebagai web unggulan di Indonesia. SINTA mengagregrasi berbagai sumber publikasi baik nasional maupun internasional, salah satu publikasi internasional yang bereputasi adalah Wos. Web ini menyediakan akses yang cepat, mudah, serta menyeluruh untuk menilai kinerja peneliti, jurnal elektronik, serta institusi di Indonesia. Setiap tenaga pendidik, program studi, dan perguruan tinggi berupaya mengembangkan peringkat di web SINTA. Berdasarkan skor di SINTA dapat ditentukan akumulasi sejumlah artikel ilmiah yang terindeks di Scopus. Tahun 2021 mulai ditambahkan komponen kinerja oleh Wos. Web tersebut mencakup literatur ilmiah dan memuat gabungan data yang menentukan dalam tahap suatu jurnal atau institusi perguruan tinggi pada lingkup publikasi ilmiah berdasarkan hubungan sitasi yang mempublikasikan sebuah jurnal peneliti dari suatu lembaga di Indonesia.

Indeksasi Wos didalam SINTA belum dikategorikan berdasarkan 5 bidang ilmu yang menjadi klasifikasi utama dalam perankingan perguruan tinggi didunia. Bidang ilmu yang dimaksud adalah *Art & Humanities, Engineering & Technology, Life Sciences & Medicine, Natural Sciences, Social Sciences & Management*. Klasifikasi tersebut berdasarkan pemetaan data ke dalam satu atau beberapa kelas. Terdapat berbagai cara atau teknik klasifikasi seperti *K-Nearest Neighbor*.

K-Nearest Neighbor atau KNN yaitu metode untuk mengklasifikasi data menurut jarak terpendek terhadap objek data. Penentuan nilai K di Algoritma ini didasarkan data diperoleh. Nilai K tertinggi dapat mengurangi efek noise saat mengklasifikasikan data, serta dapat membuat batas antara setiap klasifikasi. Metodologi KNN digunakan sebagai salah satu metode data mining khususnya klasifikasi, dan berfungsi sekaligus sebagai metode penelitian. Peneliti mengumpulkan data artikel terindeks secara Wos melalui *database* didalam SINTA.

Dengan demikian, dengan perkembangan sistem informasi di Indonesia saat ini melalui data mining serta menggunakan metode KNN membawa kontribusi cukup besar bagi teknologi web SINTA. Tujuan penelitian ini yakni merancang sebuah sistem yang dapat mengklasifikasikan data WoS yang terindeks pada SINTA sesuai 5 bidang ilmu menggunakan KNN. Keunikan penelitian ini mampu memetakkan atau klasifikasi data menjadi beberapa kelas yang didefinisikan sebelumnya. Hasil penelitian ini diharapkan memberi manfaat untuk pembaca mengenai saat melakukan pencarian artikel sesuai 5 bidang ilmu secara mudah dan cepat.

1.2 Perumusan Masalah

Bagaimana mengklasifikasikan koleksi data WoS di SINTA ke dalam 5 bidang ilmu menggunakan metode *K-Nearest Neighbor*.

1.3 Pembatasan Masalah

Berdasarkan perumusan masalah yang dikemukakan diatas, maka batasan masalah dalam penelitian ini sebagai berikut:

1. Klasifikasi data dikhususkan di dalam WoS yang terindeks SINTA.
2. Judul artikel yang di klasifikasi berbahasa inggris.
3. Data pada penelitian menggunakan judul publikasi.

1.4 Tujuan

Adapun tujuan dari tugas akhir yakni merancang sebuah sistem yang mampu mengklasifikasikan data WoS yang terindeks pada SINTA sesuai lima bidang ilmu menggunakan metode *K-Nearest Neighbor*.

1.5 Manfaat

Manfaat yang diharapkan dalam penelitian ini adalah sebagai berikut :

1. Memberi manfaat untuk pembaca mengenai saat melakukan pencarian artikel sesuai lima bidang ilmu secara mudah dan cepat.
2. Implementasi metode KNN, diharapkan mampu mengklasifikasikan data WoS secara komputerisasi yang terindeks pada SINTA sesuai lima bidang ilmu.

1.6 Sistematika Penulisan

Pada laporan penelitian ini, sistematika penulisan yang digunakan penulis dalam penulisan laporan tugas akhir ini adalah dapat dijelaskan sebagai berikut:

BAB I PENDAHULUAN

Pada bab I merupakan pendahuluan, penulis dapat menguraikan latar belakang penentuan judul, rumusan masalah, batasan masalah, tujuan penelitian, manfaat dari penelitian, dan sistematika penulisan tugas akhir.

BAB II TINJAUAN PUSTAKA DAN DASAR TEORI

Pada bab II mengemukakan sebuah penelitian-penelitian sebelumnya serta dasar teori yang mendasari pembahasan secara terperinci berfungsi mendukung penulis untuk memahami bagaimana konsep kerja pada klasifikasi bidang ilmu pada publikasi terindeks WoS dengan metode *K-Nearest Neigbor*.

BAB III METODOLOGI PENELITIAN

Pada bab III berisikan proses tahap penelitian dimulai dengan perancangan sistem web hasil klasifikasi bidang ilmu pada publikasi terindeks WoS menggunakan metode *K-Nearest Neigbor*.

BAB IV HASIL DAN ANALISIS PENELITIAN

Pada bab IV, penulis memaparkan hasil penelitian yaitu hasil pengujian sistem web dengan beberapa data training dan nilai K.

BAB V KESIMPULAN DAN SARAN

Pada bab V, menguraikan kesimpulan serangkaian tahapan penelitian dari awal sampai akhir berdasarkan analisis yang dilakukan serta saran untuk dikembangkan dalam penelitian selanjutnya.

BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Pada penyusunan tugas akhir penulis mengumpulkan informasi dari penelitian sebelumnya. Adapun beberapa penelitian terdahulu merupakan sebagai bahan referensi dan pertimbangan dalam penelitian ini yakni:

Penelitian sebelumnya dengan mengklasifikasikan artikel judul dan abstrak pada artikel menggunakan KNN dengan Metrik *Cosine Similarity*. Dalam penelitian ini terdapat 9 jenis pengujian skenario, terdiri dari beberapa urutan dan cara berlainan. Algoritma klasifikasi dipengaruhi berbagai faktor seperti gabungan teknik *preprocessing* beserta urutannya dengan menghasilkan hasil terbaik dan terburuk. Skenario 9 merupakan hasil klasifikasi tertinggi. Nilai akurasi, presisi, dan *recall* adalah 72.92%, 73.36%, dan 72.92%. Sedangkan hasil terendah pada penelitian tersebut yakni skenario 6, dengan nilai akurasi, *recall* beserta presisi yaitu 68.05%, 68.05%, dan 69.98%. Tahapan skenario 6 dari tahap *preprocessing* berisi *tokenizing, case folding, stemming*. Kemudian, ditransformasi dalam SMOTE atau (*Synthetic Minority Over-Sampling Technique)* dan VSM (*Vector Space Model*) (Ma'rifah dkk., 2020).

Penelitian berikutnya dengan analisis sentimen data twitter dengan metode KNN, *Decision Tree*, dan *Naïve Bayes* terhadap layanan BPJS yang nilai akurasi dalam metode KNN 96.01%. Demikian nilai presisi untuk pred. *positive* adalah 0.00%, pred. *negative* 52.17%, dan pred. *neutral* adalah 97.27%. Akurasi pada metode *Decision Tree* bernilai 96.13%, nilai presisi untuk pred. *negative* 55.00%, dan pred. *positive* adalah 0.00%, serta pred. *neutral* adalah 97.28%. Terakhir metode *Naïve Bayes* dengan nilai akurasi 89.14%, nilai presisi untuk pred. *positive*, pred. *negative*, dan pred. *neutral* yakni16.67%, 1.64%, dan 98.40%.(Puspita & Widodo, 2021)

Penelitian tentang Aplikasi untuk mengimplementasikan dan memverifikasi metode K-Nearest Neighbor untuk menyeleksi karyawan baru, sistem ini bekerja dengan baik di Google Chrome dan Microsoft Edge. Berdasarkan hasil perhitungan algoritma KNN menggunakan K = 7 pada metode *Euclidean Distance* diperoleh nilai *accuracy, precision* dan *recall* masing-masing

adalah 91 %, 87%, dan 100%. Berikut hasil penelitian metode, yaitu dengan perhitungan manual menggunakan Microsoft Excel menghasilkan nilai akurasi 100% dan persentase *error* sebesar 0%. Sedangkan hasil pengujian program, mengasilkan persentase akurasi 80% dan persentase *error* sebesar 20%. Dan hasil pengujian *user* pada 2 *user* memperoleh nilai 8 untuk penilaian terbaik dan penilaian cukup dengan nilai 4 (Rahmat Dian Nugraha dkk., 2020).

Penelitian selanjutnya, pada permasalahan alumni STIKOM Bali dalam mengetahui waktu untuk mendapatkan pekerjaan dengan cara mengklasifikasikan kasus yang ada, tolok ukur dalam penelitian ini yakni jenis kelamin, IPK, dan masa studi. Perhitungan MAPE serta akurasi penelitian tersebut menetapkan dua metode yaitu metode *Naïve bayes* serta metode KNN untuk nilai K yaitu 3,5,7, dan 9. Untuk dapat menghasilkan tingkatan akurasi secara relevan dengan kedua metode, peningkatan ini mempengaruhi peralihan sebuah data atribut rentang waktu. Uji coba menggunakan jumlah *class* percobaan pertama yakni lima, sedangakan pengujian kedua hanya 2 *class*. Alhasil adanya peningkatan diatas, disimpulkan metode diatas kurang baik dengan beberapa jumlah *class*. Data *training* dalam penelitian ini berjumlah 1335 dan data *testing* 334, Naïve Bayes memperoleh nilai akurasi dan perhitungan MAPE terbaik yaitu 83.83% dan 16.17% sedangkan KNN dengan nilai K 9 adalah nilai terbaik. Nilai akurasi adalah 82.34% dan MAPE 17.66%. Berdasarkan nilai MAPE, KNN dengan nilai K= 9 dan *Naïve Bayes* dapat diartikan bahwa kedua metode pada penelitian ini, Naïve Bayes sedikit lebih unggul (Maricar & Dian Pramana, 2019).

Selanjutnya, metode *Modified K-Nearest Neighbor* untuk mengklasifikasikan informasi yang disebarkan di seluruh sistem sebagai hoaks. Data terlebih dahulu diproses melalui *preprocessing* sehingga diperoleh kata-kata yang paling sering muncul dari setiap paragraf. Dengan *Term Frequency Inverse Document Frequency* atau (TF-IDF), selanjutnya dilakukan pembobotan setiap data latih dan data uji. Setelah itu dilakukan, pembobotan dari setiap data diperoleh, dilakukan perbandingan antara dokumen versi latih dan versi uji dengan menggunakan *Cosine Similarity*. Langkah terakhir adalah mengklasifikasikan dokumen dengan menggunakan bobot dan mengukurnya. Hasil klasifikasi 466 dokumen latih serta

65 dokumen uji. Dapat diperoleh nilai akurasi, presisi, dan *recall* sebesar 92.30%, 93.75% dan 90.90%. Menentukan hasil klasifikasi dari perhitungan sistem bergantung pada kualitas kata setiap dokumen. Pengujian yang tidak akurat karena data latih yang diperoleh dalam sistem tidak lengkap. Sehingga menimbulkan faktor seperti rendahnya kemiripan antar dokumen (Rozi & Sulistyawati, 2019).

Berikut penelitian selanjutnya, klasifikasi pada berita olahraga melalui metode pembobotan dan pemeringkatan BM25 dengan klasifikasi metode KNN berisi tahapan perencanaan hingga tahap pengujian yang dapat terselesaikan. Metode pembobotan, pemeringkatan BM25 serta klasifikasi metode *K-Nearest Neighbor* pada penelitian ini dapat diaplikasikan. Dari proses *pre-processing* sampai tahap penyortiran nilai BM25. Sedangkan pengujian ini menggunakan Presisi, *Recall*, dan F-Measure untuk mengevaluasi sistem. Penelitian tersebut, menyimpulkan bahwa hasil tertinggi diperoleh pada saat nilai K yakni 20, dengan nilai presisi 0.921577, nilai *recall* yakni 0.914286 serta *f-measure* adalah 0,917917. Sedangkan skor tes terendah diperoleh nilai K 200, dengan nilai presisi, *recall*, dan *f-measure* adalah 0,732889, 0,714286, dan 0,699871 (Septrinas Enggar, Indriati, 2019).

Berdasarkan penelitian ini dapat diketahui bahwa penerapan pada seleksi fitur dapat mencapai tingkatan nilai akurasi, presisi, r*ecall*, *F1-Score*, dan akurasi dalam metode *K-Nearest Neighbor* untuk mengklasifikasikan daun dan seleksi fitur chi-square lebih baik daripada pemilihan fungsi *Gini Index*. Pada percobaan seleksi fitur *chi-square*, nilai parameter threshold kombinasi yaitu 1% dengan parameter nilai k=6 merupakan kombinasi terpilih sebagai model terbaik. Batas ini memilih sekitar 31 fitur dengan nilai *chi-square* terbesar. Saat menguji data baru, model KNearest Neighbor dengan pemilihan fitur chi-square memberikan akurasi 85%, 83,3%, 83,3%, 88,2%, dan 92,3%, skor F1, dan 92,3%. Pada eksperimen pemilihan fitur Gini *Index*, kombinasi three1 dan nilai parameter k=4 merupakan kombinasi terpilih model terbaik. Batas ini memilih sekitar 31 fitur dengan nilai indeks Gini tertinggi. Saat menguji data baru, model *K-Nearest Neighbor* dengan pemilihan fitur Gini *Index* memberikan kinerja presisi, pemulihan, skor F1, dan presisi, yaitu masing-masing 81,2%, 80,3%, dan 81,6%,dan 86,6 (Istighfarizky dkk., 2022).

Penelitian pada klasifikasi judul berita Hoax dengan penerapan algoritma KNN memberikan hasil akurasi tertinggi sebesar 93,33%, presisi 100%, *recall* 80% serta skor f1 88,89%. Dan dapat ditarik kesimpulan metode diatas yaitu algoritma *K-Nearest Neighbor* berfungsi klasifikasi *headline* berita. Skenario untuk dalam proses klasifikasi penelitian tersebut yakni Skenario 1 (90:10), Skenario 2 (80:20) dan Skenario 3 adalah (70:30). Penggunaan Skenario dalam penelitian diatas terbukti mempengaruhi hasil akurasi penelitian (Hendriyanto & Sari, 2022).

2.2 Dasar Teori

2.2.1 *Science and Technology Index* (SINTA)

Gambar 2. 1 Logo *Web* SINTA

SINTA adalah suatu web yang menyediakan akses untuk kutipan ilmiah serta keahlian peneliti. Sistem informasi tersebut memberikan fungsi yang mudah dan cepat dalam mengetahui peforma peneliti dalam jurnal perguruan tinggi di Indonesia. SINTA merupakan web publikasi yang indeks meliputi ilmu pengetahuan dan teknologi yang meliputi kekayaan intelektual berserta dampaknya (sitasi), pengabdian kepada masyarakat, dan kepakaran di Indonesia. SINTA mampu mendukung institusi akademik, jurnal, serta peforma peneliti di Indonesia (SIADNYANI, 2018).

Sistem informasi tersebut menawarkan informasi yang berhubungan dengan kriteria institusi pendidikan, yang berkolaborasi dengan direktori pakar Indonesia untuk menganalisis penelitian terbaru. SINTA bertujuan dalam mempermudah publikasi dari karangan tenaga pendidik perguruan tinggi, peneliti, dan institusi yang ada di Indonesia. Sehingga, dapat diketahui hasil prestasi kerjanya (Sepdela, 2018).

2.2.2 *Web of Science* (WoS)

Gambar 2. 2 Logo *Web of Science*

WoS merupakan sebuah web pengindeksan kutipan ilmiah berbasis online. WoS menyediakan pencarian kutipan yang lengkap secara menyeluruh. Layanan ini menyediakan akses ke beberapa *database* referensi penelitian seperti SINTA, dan memungkinkan dalam penelitian mendalam untuk berbagai bidang akademis atau ilmiah. WoS sebagai web penelitian keseluruhan yang memungkinkan pengguna untuk memperoleh, menganalisis, dan menyebarkan informasi *database* secara tepat. WoS menggunakan berbagai kemampuan pencarian dan analisis. Pengindeksan kutipan digunakan yang ditingkatkan dengan kemampuan untuk mencari hasil lintas bidang ilmu, pengaruh dampak, sejarah, dan metodologi suatu ide. Teknologi ini menunjukkan kekurangan dengan metode pencarian kata kunci saja.

2.2.3 Data Mining

Proses menganalisis data untuk memperoleh data yang tidak diketahui sebuah pengertian dari Data Mining. Selanjutnya, dapat diartikan mempersingkat data menggunakan teknik berbeda dan mudah dipahami bagi peneliti (Yuli Mardi, 2019). Ada berbagai kelompok teknik dalam data mining. Salah satunya yaitu Model prediksi yang terhubung dengan penyusunan untuk pemetaan pada setiap variabel hingga tujuannya. Kemudian, model berpengaruh untuk menunjukkan nilai target pada himpunan baru yang diperoleh. Pada data mining dalam model prediksi ada 2 yakni, regresi dan klasifikasi. Dalam penelitian ini menggunakan jenis klasifikasi. Klasifikasi berguna pada variabel target diskret, misalnya ingin mengklasifikasi bidang ilmu pada publikasi terindeks *Web of Science*. kasus

tersebut merupakan termasuk dalam jenis klasifikasi, karena pengkelompokkan sesuai bidang ilmu adalah variabel target diskret.

Data mining suatu istilah untuk menggambarkan penemuan informasi dari basis data. Proses teknik statistik, kecerdasan buatan, matematika, dan pembelajaran mesin berguna untuk mempelajari, mengidentifikasi informasi basis data yang relevan pada data mining.

Data dengan tujuan aplikasi data mining harus sesuai model. Hasil diharapkan dapat digunakan untuk meemperoleh keputusan dan melakukan kajian yang diperlukan. Proses KDD atau (*Knowledge Discovery in Databases)* merupakan bagian proses. KDD yaitu menemukan informasi yang mudah serta dapat dipahami dari data terkait. Proses KDD menguraikan hasil data yang di peroleh dan menggabungkan pada bidang ilmu lain (Sri, 2019).

2.2.4 Klasifikasi

Suatu proses perancangan atau model berfungsi mendeskripsikan dan memisahkan kelas atau kerangka yang menaksir kelas dari objek berlabel tidak diketahui. Pada pengertian diatas, dimana data uji untuk memperkirakan keakuratan, menganalisis data, kemudian dipresentasikan dalam bentuk aturan klasifikasi (Rahmat Dian Nugraha dkk., 2020). Klasifikasi mengevaluasi objek data dan menempatkan data dengan kategori tertentu dari jumlah dalam kategori yang diketahui.

Klasifikasi membuat pola berdasarkan data pelatihan. Setelah itu, menggunakan pola untuk mengklasifikasikan data. Suatu sistem diinginkan dapat klasifikasi sesuai data, sehingga sistem klasifikasi juga dapat mengukur kinerja. Klasifikasi data mining dibagi tiga antara lain yakni, *supervised*, *unsupervised*, dan *semi-supervised*. Pada *supervised learning*, klasifikasi berproses dalam suatu kumpulan data berlabel atau memiliki kelas yang diketahui (Sri, 2019).

Dalam *supervised learning*, pengidentifikasi atau kategori data tidak ketahui, data dikelompokkan berdasarkan kesamaannya. Klasifikasi merupakan t *supervised learning*. Tahapan klasifikasi dibedakan menjadi dua tahapan:

1. Tahapan Membangun Model

   Tahapan ini adalah suatu langkah membuat model klasifikasi berdasarkan data kelas yang diberikan. Data sampel diperoleh disebut data latih atau data uji.

2. Tahapan Menggunakan Model Klasifikasi

   Tahapan ini merupakan model yang menerapka data kelas yang tidak diketahui. Proses menerapkan model ini yaitu memprediksi label kelas suatu data dalam kumpulan data uji (Data *Testing*).

Secara umum, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi. Pada penelitian ini penulis mengklasifikan publikasi yang terindeks WoS sesuai dengan 5 bidang ilmu secara komputerisasi.

2.2.5 *K-Nearest Neighbor* (KNN)

Algoritma klasifikasi menggunakan hasil model yang relatif terhadap data masukan dan data sebelumnya. Algoritma dapat dikenali dari data yang diklasifikasikan dengan benar, atau seberapa akurat data model dapat memprediksi data kelas yang diketahui. A lgoritma dan teknik yang digunakan dalam data mining salah satunya yakni Algoritma *K-Nearest Neighbor* (KNN). Algoritma KNN adalah bagian *supervised learning*, sebagai hasil nilai *instance query* menurut data terbanyak pada kategori tersedia. Klasifikasi menghasilkan kelas yang sering muncul. Data yang diambil adalah rata-rata terhitung dari awal, dan langkah selanjutnya menghitung kuadrat menggunakan *Euclidean Distance* (Jarak *Euclidean*) terhadap data uji. Nilai K bertujuan menetapkan jumlah tetangga terdekat pada metode KNN dari suatu data latih. Jumlah nilai K dipilih berdasarkan keakuratan hasil. Kemudian, untuk mendapatkan nilai K perlu membandingkan beberapa nilai K. Misalnya yaitu nilai K=3 sampai nilai K yang ditentukan. Menentukan nilai K adalah cara untuk menghitung Algoritma KNN, menghitung kuadrat jarak *Euclidean* masing-masing data testing pada data training, data hasil perhitungan jarak *Euclidean* diurutkan dari terkecil, mengumpulkan label paling banyak (klasifikasi KNN). Label mayoritas yaitu kelompok klasifikasi dari data training.

Ada banyak cara dalam algoritma KNN untuk mengukur kedekatan antar data, termasuk menggunakan jarak *Euclidean*. Jarak *Euclidean* merupakan metode

yang berfungsi untuk menilai jarak antar data. Jarak tersebut untuk mengukur kedekatan nilai interpretasi jarak antara dua kata. Berikut ini persamaan untuk jarak *Euclidean* (Kasanah dkk., 2019):

$$d(P,Q) = \sqrt{\sum_{i=1}^{n}(P_{i2}-Q_{i1})^2} \quad (1)$$

Dimana

d(P,Q) : Jarak Euclidean

n : Jumlah atribut.

P_{i2} : Inputan data uji.

Q_{i1} : Inputan data sampel.

i : Atribut

Dengan adanya Algoritma KNN dapat mempermudah penulis dalam mengklasifikasi bidang ilmu pada publikasi Terindeks Web of Science.

2.2.6 Evaluasi Mesin Klasifikasi

Evaluasi mesin klasifikasi berlandaskan dari uji coba data yang sesuai dan tidak sesuai. Validasi tersebut diharapkan untuk menyakinkan suatu bentuk model yang terbaik dari sebuah hasil klasifikasi penelitian. Hasil evaluasi mesin klasifikasi dalam penelitian ini dengan matriks konfusi. Matrik konfusi (*Confusion matrix*) adalah informasi yang dapat diprediksi oleh sistem klasifikasi tentang hasil klasifikasi yang sebenarnya.

Tabel 2. 1 Tabel *Confusion Matrix*

		Nilai Sebenarnya	
		TRUE	**FALSE**
Nilai Prediksi	TRUE	TP (*True Positive*)	FP (*False Positive*)
	FALSE	FN (*False Negative*)	TN (*True Negative*)

untuk mengukur kemampuan sistem klasifikasi yang dibangun ada tiga nilai, yaitu akurasi, presisi, dan *recall*. Nilai presisi suatu nilai akurasi atau nilai ketepatan sistem diantara data yang diberikan oleh sistem untuk menampilkan dengan benar data kelas negatif atau positif.(Azhari dkk., 2021)

Nilai yang memberitahukan presentase keberhasilan atau memperoleh kembali sebuah informasi secara benar tentang data yang kelas *negatif* ataupun

konten teks *positif* disebut nilai *recall*. Nilai presisi dan *recall* dapat diketahui pada rumus persamaan berikut (Kasanah dkk., 2019):

$$\text{Rumus } Presisi = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Rumus } Recall = \frac{TP}{TP + FN} \quad (3)$$

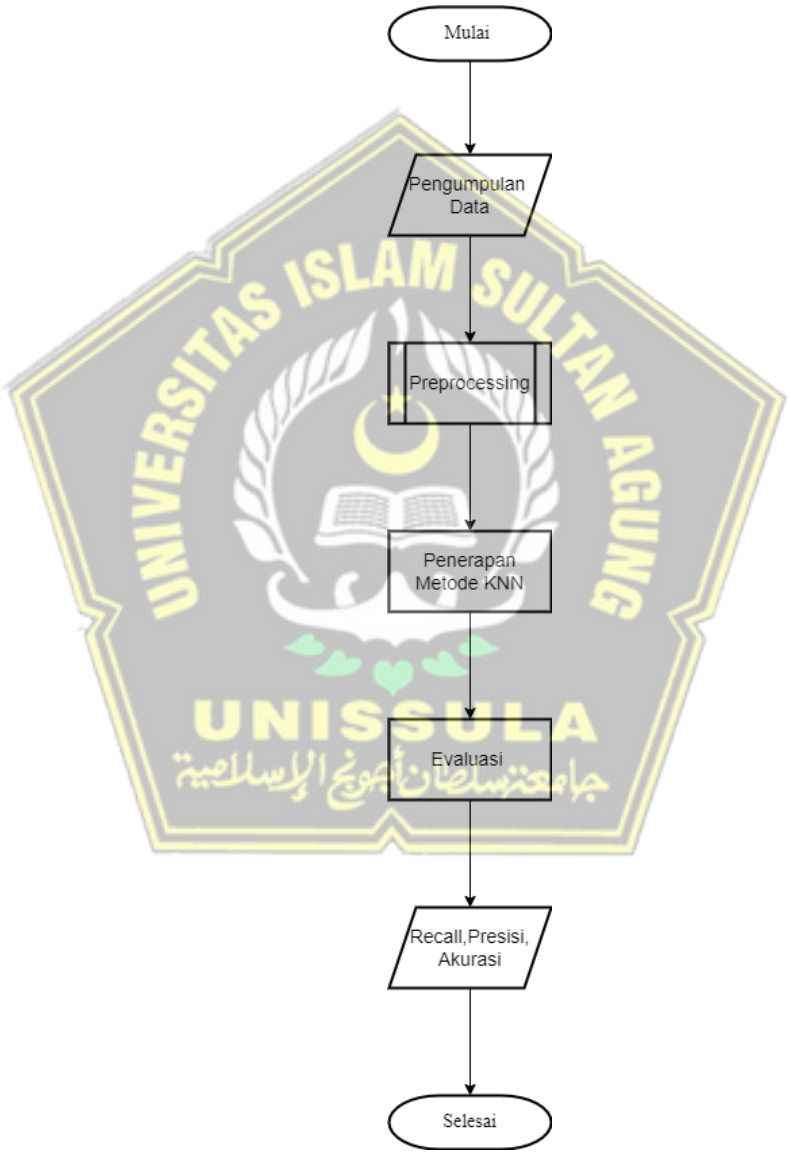
Sedangkan nilai akurasi yakni nilai rasio data bernilai benar yang terdeteksi pada data pengujian. Dengan kata lain, akurasi adalah nilai yang menampilkan tingkatan nilai sebuah prediksi. Nilai akurasi dirumuskan dengan persamaan berikut:

$$\text{Rumus } Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

BAB III
METODE PENELITIAN

3.1 Tahapan Penelitian

Pada tahap penelitian, algoritma atau metode yang pada penelitian yaitu *K-Nearest Neighbor*. Metode ini mengklasifikasi kumpulan data WoS di SINTA ke dalam 5 bidang ilmu. Adapun langkah tahapan yang digunakan dalam penelitian ini sebagai berikut:



Gambar 3. 1 *Flowchart* Tahapan Penelitian

Pada 3.1 merupakan *flowchart* tahapan dari penelitian ini, terdiri atas pengumpulan data dari WoS pada *database* SINTA. Selanjutnya, dilakukan tahapan preprocessing, setelah melalui proses

tersebut dihasilkan sebuah data. Langkah berikut, penerapan metode yaitu KNN, tahapan yang terakhir evaluasi, didalam evaluasi mengukur data dengan menggunakan rumus akurasi, presisi, dan *recall* untuk mengetahui keakuratan.

1. **Pengumpulan Data**

   Pengumpulan data adalah tahapan proses, penelitian berlanjut hingga diperoleh sebuah data sesuai dengan peneliti yang telah diidentifikasinya. Data digunakan berdasarkan tujuan penelitian. Dengan teknik yang tepat, dapat memperoleh strategi dan prosedur akan digunakan. Pengumpulan data penelitian ini diambil dari WoS pada *database* SINTA yang belum sesuai dengan lima bidang ilmu. Pada penelitian ini menggunakan *dataset* untuk eksperimen berjumlah 1000 judul publikasi yang mencakup lima bidang ilmu yang telah dikemukakan diatas.

2. ***Preprocessing***

   Preprocessing adalah sebuah langkah atau proses awalan data mining saat mengubah data yang diperoleh dari berbagai sumber untuk menjadi data yang bersih, kemudian data tersebut untuk mengolah data berkelanjutan.

3. **Penerapan Metode *K-Nearest Neighbor* (KNN)**

   Dalam tahapan ini, setelah data diolah melalui tahap *Preprocessing*, menghasilkan sebuah data bersih yang selanjutnya dilakukan tahap klasifikasi dengan menggunakan metode KNN. Penerapan metode dalam penelitian ini diharapkan mendapatkan data yang diharapkan.

4. **Evaluasi**

   Pada tahap ini dilakukan evaluasi metode *K-Nearest Neighbor* dengan mencakup hasil eksperimen berdasarkan mengukur performa dari recall, presisi dan akurasi. Setelah adanya beberapa pengujian penelitian akan dipilih dari data yang terbaik.

3.1.1 Pengumpulan Data WoS

Data dari WoS yang telah dikumpulkan dalam penambangan tugas akhir ini adalah sebanyak 1000 data, yang mana data tersebut didapatkan dari database SINTA. Adapun beberapa sampel dari database SINTA yang mana telah didapatkan oleh penulis sebagai berikut:

Tabel 3. 1 Sampel Data

Label	Judul
Art & Humanities	The Impact of Comprehension Instruction on Students' Reading Comprehension with Different Ability Grouping And Self-Efficacy
Engineering & Technology	Potential Analysis of Ex-Coal Mining Land As Pumped Storage Hydro Powerplant in Kutai Kartanegara East Kalimantan
Life Science & Medicine	Anti-Vibrio from Ethyl Acetate Extract of Sponge-Associated Fungus Trichoderma longibrachiatum
Social Sciences & Management	The effect of Liquidity and Leverage on Financial Distress (study on idx food and beverage sub-sector manufacturing companies for the 2015-2020 period).
Natural Sciences	The relationship between the level of corruption and economic growth in indonesia: An investigation using supply chain strategy and bounds test

3.1.2 Preprocessing

Tahap *preprocessing* dalam penelitian ini yaitu proses pengkodean data pada judul artikel di SINTA yang terindeks oleh WoS . Berikut beberapa tahapan pada tahapan data *preprocessing* adalah:

```
// pipeline untuk preprocessing dan klasifikasi
$estimator = new PersistentModel(
    new Pipeline([
        //PREPROCESSING
        new RegexFilter([
            RegexFilter::EXTRA_WHITESPACE,
            RegexFilter::EXTRA_WORDS
        ]), //text cleaning
        new MultibyteTextNormalizer(), //case folding - lower case
        new StopWordFilter($stopwords_en), //stopwords
        new WordCountVectorizer(), //tokenizer (hapus untuk Bayes)
```

Gambar 3. 2 *Syntax* Tahapan *Preprocessing*

Pada gambar 3.2 merupakan *code* yang digunakan pada proses tahapan *preprocessing* diantara lain yaitu proses data *cleaning, case folding, tokenizing, stopword,* dan terakhir *stemming.*

1) *Cleaning* merupakan sebuah langkah proses menghilangkan karakter tanda baca pada teks seperti angka dan simbol.

Tabel 3. 2 Perubahan data sebelum dan sesudah *cleaning*.

Label	Judul (Sebelum *Cleaning*)	Judul (Sesudah *Cleaning*)
Engineering & Technology	Potential Analysis of Ex-Coal Mining Land As Pumped Storage Hydro Powerplant in Kutai Kartanegara, East Kalimantan	Potential Analysis of Ex Coal Mining Land As Pumped Storage Hydro Powerplant in Kutai Kartanegara East Kalimantan

```
new RegexFilter([
    RegexFilter::EXTRA_WHITESPACE,
    RegexFilter::EXTRA_WORDS
]), //text cleaning
```

Gambar 3. 3 *Syntax* proses *Cleaning*.

Pada gambar 3.3 merupakan codingan dari proses cleaning dimana memiliki fungsi untuk menghilangkan karakter tanda baca pada teks seperti angka dan symbol yang tidak sesuai.

```
new RegexFilter([
```

Gambar 3. 4 *Syntax RegexFilter*.

Regexfilter merupakan fungsi menyaring fitur teks pada kumpulan data dengan mencocokkan dan menghapus pola dari daftar.

```
RegexFilter::EXTRA_WHITESPACE,
```

Gambar 3. 5 *Syntax RegexFilter EXTRA_WHITESPACE*.

Fungsi *RegexFilter* dengan *EXTRA_WHITESPACE* diartikan sebagai metode untuk mendeteksi suatu pola dari string dengan mencocokan karakter spasi kosong yang diulang secara berurutan. Pada sistem ini fungsi tersebut dapat mendeteksi karakter, spasi yang kosong,

```
RegexFilter::EXTRA_WORDS
```

Gambar 3. 6 *Syntax RegexFilter EXTRA_WORDS*.

Fungsi *RegexFilter* dengan *EXTRA_WORDS* dapat diartikan sebagai metode untuk mendeteksi suatu pola dari string dengan mencocokan kata-kata yang diulang secara berurutan. Pada sistem ini fungsi tersebut mendeteksi kata-kata pada publikasi terindeks WoS.

2) *Case Folding* merupakan sebuah proses merubah seluruh data menjadi huruf kecil.

Tabel 3. 3 Perubahan data sebelum dan sesudah *case folding.*

Label	Judul (Sebelum *Case Folding)*	Judul (Sesudah *Case Folding)*
Engineering & Technology	Potential Analysis of Ex Coal Mining Land As Pumped Storage Hydro Powerplant in Kutai Kartanegara East Kalimantan	potential analysis of ex coal mining land as pumped storage hydro powerplant in kutai kartanegara east kalimantan

```
new MultibyteTextNormalizer(), //case folding - lower case
```

Gambar 3. 7 *Syntax MultibyTextNormalizer.*

Fungsi *MultibyTextNormalizer* adalah mengubah karakter dalam semua string menjadi case yang sama. Dalam hal sistem ini fungsi memiliki peran yaitu mengubah kata-kata judul publikasi ke dalam huruf kecil (*lower case*).

3) *Tokenizing* adalah proses memecah serangkaian karakter dari teks menjadi kata-kata.

Tabel 3. 4 Perubahan data sebelum dan sesudah *tokenizing.*

Label	Judul (Sebelum *tokenizing)*	Judul (Sesudah *tokenizing)*
Engineering & Technology	Potential Analysis of Ex Coal Mining Land As Pumped Storage Hydro Powerplant in Kutai Kartanegara East Kalimantan	‘potential’ ‘analysis’ ‘of’ ‘ex’ ‘coal’ ‘mining’ ‘land’ ‘as’ ‘pumped’ ‘storage’ ‘hydro’ ‘powerplant’ ‘in’ ‘kutai’ ‘kartanegara’ ‘east’ ‘kalimantan’.

```
new WordCountVectorizer(), //tokenizer
```

Gambar 3. *8 Syntax WordCountVectorizer.*

Fungsi *WordCountVectorizer* adalah fungsi untuk membangun kosakata dari sampel data. Pada sistem ini fungsi tersebut berperan untuk mengubah judul publikasi menjadi serangkaian kata-kata.

4) *Stopword* adalah proses dimana kata sambung dapat dihilangkan.

Tabel 3. 5 Perubahan data sebelum dan sesudah *Stopword.*

Label	Judul (Sebelum *Stopword*)	Judul (Sesudah *Stopword*)
Engineering & Technology	'potential' 'analysis' 'of' 'ex' 'coal' 'mining' 'land' 'as' 'pumped' 'storage' 'hydro' 'powerplant' 'in' 'kutai' 'kartanegara' 'east' 'kalimantan'.	potential analysis coal mining land pumped storage hydro powerplant kutai kartanegara east Kalimantan.

```
new StopWordFilter($stopwords_en), //stopwords
```

Gambar 3. 9 *Syntax StopWordFilter*.

Fungsi *StopWordFilter* yaitu menghapus kata yang ditentukan pengguna dari kolom fitur kategori. Pada fungsi tersebut dalam sistem ini berperan menghilangkan kata sambung pada judul publikasi terindeks WoS.

5) *Stemming* adalah proses pencarian suatu kata dasar pada setiap kata.

Tabel 3. 6 Perubahan data sebelum dan sesudah *Stemming.*

Label	Judul (Sebelum *Stemming*)	Judul (Sesudah *Stemming*)
Engineering & Technology	potential analysis coal mining land pumped storage hydro powerplant kutai kartanegara east kalimantan.	potency analyse coal mine land pump storage hydro powerplant kutai kartanegara east Kalimantan.

```php
// fungsi untuk stemmer bahasa inggris
$stemmeringgris = function (&$sample, $offset, $context) {
    $stemmer_en = new WordStemmer('english');
    $words = new Word();
    $arrwords = [];
    foreach ($sample as $s) {
        if ($s !== null) {
            $array_words = $words->tokenize($s);

            foreach ($array_words as $aw) {
                if ($aw !== null) {
                    $stemmres = $stemmer_en->tokenize($aw);
                    $arrwords[] = $stemmres[0];
                }
            }
            $arrimploded = implode(" ", $arrwords);
        }
        $sample[0] = $arrimploded;
    }
```

Gambar 3. 10 *Syntax* proses *stemming.*

Pada gambar 3.10 merupakan codingan proses *preprocessing* pada tahapan *stemming*. *Stemming* merupakan proses dimana pencarian suatu kata dasar.

3.1.3 Penerapan Metode KNN

Setelah data diolah melalui tahap *preprocessing*, menghasilkan sebuah data bersih yang selanjutnya dilakukan tahapan klasifikasi menggunakan metode KNN. Penerapan metode pada penelitian ini diharapkan mendapatkan data yang telah diharapkan.

```php
        // ALGORITMA KLASIFIKASI
    ], new KNearestNeighbors(50)),
```

Gambar 3. 11 *Syntax* penerapan KNN.

Pada Gambar 3.11 Merupakan code yang digunakan implementasi penerapan metode yaitu menggunakan metode KNN dengan menyertakan nilai K.

```php
        // ALGORITMA KLASIFIKASI
    ], new KNearestNeighbors(50)),
```

Gambar 3. 12 *Syntax* Algoritma KNN.

Algoritma Klasifikasi pada sistem ini adalah algoritma KNN dimana jarak k pada sampel terdekat dari data dan dapat memprediksi label kelas.

Evaluasi hasil akhir pada penelitian ini menggunakan *confusion metrix* terhadap hasil akurasi, presisi, *recall* dengan melakukan berbagai percobaan penetuan jumlah tetangga terdekat (k) yaitu K = 15,25,35,45, dan 55.

```
    // ALGORITMA KLASIFIKASI
], new KNearestNeighbors(15)),
```

Gambar 3. 13 *Syntax* Algoritma KNN dengan menggunakan nilai K 15.

Gambar diatas merupakan syntax algoritma KNN yang menggunakan nilai k 15 yang bertujuan untuk melakukan percobaan penelitian.

```
    // ALGORITMA KLASIFIKASI
], new KNearestNeighbors(25)),
```

Gambar 3. 14 *Syntax* Algoritma KNN dengan menggunakan nilai K 25.

Pada gambar 3.22 merupakan code algoritma KNN dengan nilai k berjumlah 25, dimana untuk percobaan pengujian.

```
    // ALGORITMA KLASIFIKASI
], new KNearestNeighbors(35)),
```

Gambar 3. 15 *Syntax* Algoritma KNN dengan menggunakan nilai K 35.

Pada gambar diatas yaitu gambar 3.23 ialah sintaks algoritma KNN yang digunakan pada klasifikasi bidang ilmu terindeks *Web of Science* dengan nilai k berjumlah 35.

```
    // ALGORITMA KLASIFIKASI
], new KNearestNeighbors(45)),
```

Gambar 3. 16 *Syntax* Algoritma KNN dengan menggunakan nilai K 45.

Pada gambar diatas yaitu gambar 3.24 adalah sebuah sintaks algoritma KNN yang digunakan pada penelitian ini yaitu sistem klasifikasi bidang ilmu terindeks *web of science* dengan nilai k berjumlah 45.

```
    // ALGORITMA KLASIFIKASI
], new KNearestNeighbors(55)),
```

Gambar 3. 17 *Syntax* Algoritma KNN dengan menggunakan nilai K 55.

Pada gambar diatas yaitu gambar 3.25 merupakan sintaks algoritma KNN yang berperan pada penelitian ini yaitu sistem klasifikasi bidang ilmu terindeks *web of science* dengan nilai k berjumlah 55.

3.1.4 Evaluasi

```
// testing menggunakan data testing dari dataset
$predictions = $estimator->predict($testing);

// validasi, confusion matrix, akurasi, dll
$report = new AggregateReport([
    'breakdown' => new MulticlassBreakdown(),
    'confussion_matrix' => new ConfusionMatrix(),
]);

$results = $report->generate($predictions, $testing->labels());

echo $results;
```

Gambar 3. 18 *Syntax* proses evaluasi.

Pada Gambar 3.14 Merupakan code yang digunakan implementasi evaluasi pada sistem ini yang berisi prediksi, akurasi, presisi, dan recall beserta perhitungan Confusion matrix.

```
// testing menggunakan data testing dari dataset
$predictions = $estimator->predict($testing);
```

Gambar 3. 19 *Syntax* Prediksi.

Fungsi dari *function predictions* untuk menghitung skor validasi, pada prediksi dari estimator beserta label yang diharapkan.

```
// validasi, confusion matrix, akurasi, dll
$report = new AggregateReport([
```

Gambar 3. 20 *Syntax* pembuat laporan.

Yang dimaksud dari *query AggregateReport* yaitu Pembuat laporan yang menggabungkan keluaran dari beberapa laporan. Jadi pada sistem fungsi tersebut berupa Akurasi, presisi, recall, dan *Confusion matrix*.

```
    'breakdown' => new MulticlassBreakdown(),
    'confussion_matrix' => new ConfusionMatrix(),
]);
```

Gambar 3. 21 *Syntax* klasifikasi.

Fungsi *MulticlassBreakdown* yaitu Laporan klasifikasi multikelas yang menghitung sejumlah metrik (Akurasi, Presisi, recall, dll.)

```
    'confussion_matrix' => new ConfusionMatrix(),
]);
```

Gambar 3. 22 *Syntax confusion matrix.*

Pada *Syntax ConfusionMatrix* adalah matriks persegi (tabel) yang memvisualisasikan positif benar, positif salah, negatif benar, dan negatif salah dari serangkaian prediksi dan label yang sesuai.

```
$results = $report->generate($predictions, $testing->labels());

echo $results;
```

Gambar 3. 23 *Syntax function generate.*

Hasil dari *function report generate* yaitu (*function* prediksi, *function testing* label). Dimana fungsi ini menampilkan sebuah prediksi data.

```
// membagi dataset menjadi data testing dan training
[$training, $testing] = $dataset->stratifiedSplit(0.9);
```

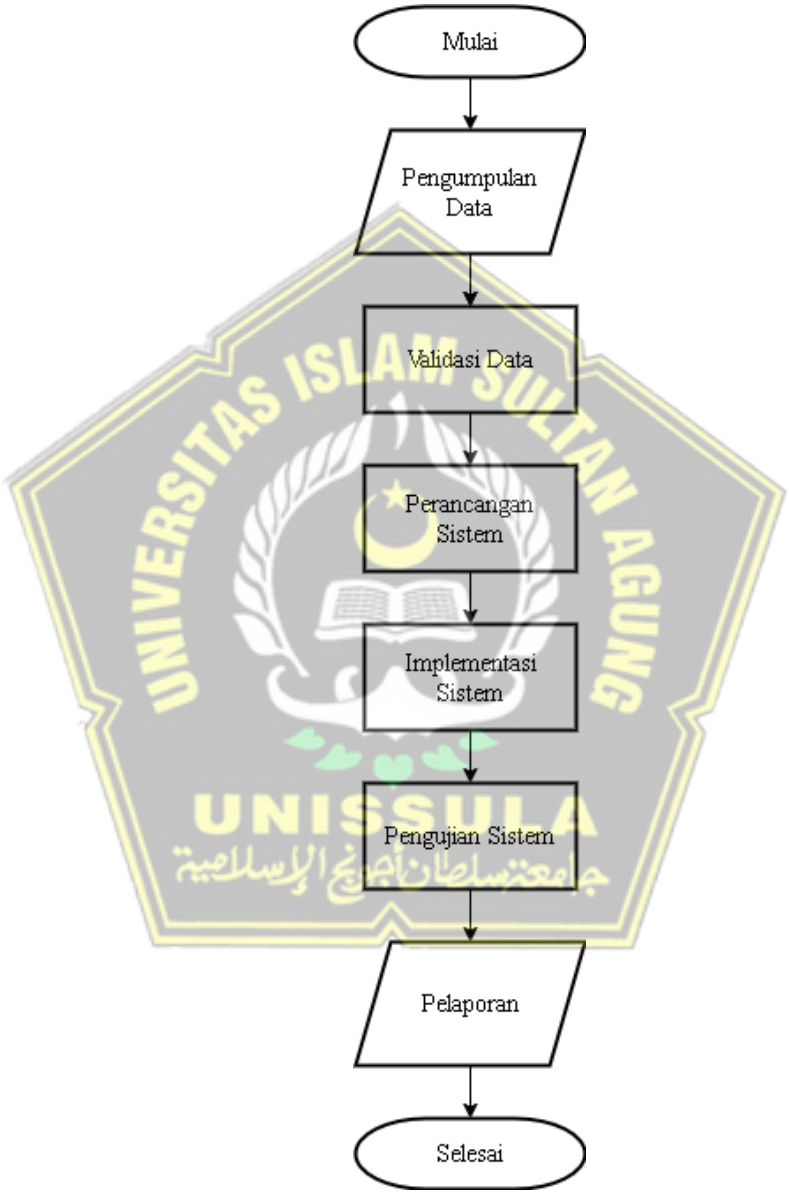
Gambar 3. 24 *Syntax Split* data.

Pada gambar diatas merupakan code yang dibuat untuk skenario data uji. Pada penelitian ini digunakan skenario data uji berupa 90%:10%, dimana data training berjumlah 900 sedangkan data testing 100 data.

3.2 Perancangan Sistem

3.2.1 Desain Sistem

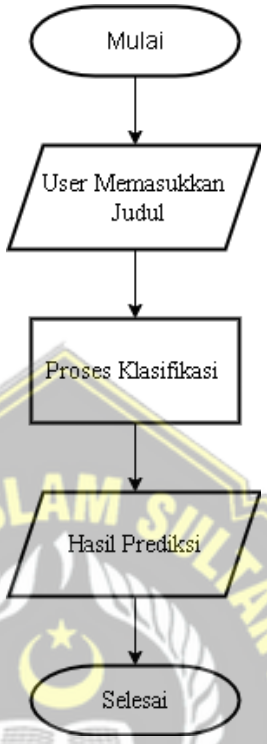
Berikut penjelasan desain sistem, akan dibuat sebuah *flowchart* yang menunjukkan alur perancangan dan sekaligus alur kerja dari sistem ini, dimana *flowchart* dari alur perancangan dapat dilihat pada gambar 3.25



Gambar 3. 25 Alur perancangan sistem.

Pada gambar 3.25 Diperlihatkan urutan dalam perancangan sistem dimana pada tahap pertama terdapat, pengumpulan data dari database SINTA, tahapan selanjutnya validasi data, kemudian jika data benar langkah selanjutnya yaitu

perancangan alur sistem dan dilanjutkan kedalam implementasi *coding* dan selanjutnya adalah pengujian sistem dan yang terakhir adalah pelaporan.



Gambar 3. 26 Flowchart Alur Sistem.

Pada gambar 3.26 adalah *flowchart* dari sistem ini nantinya, alur dari penggunaan sistem ini nantinya akan diawali dengan user memulai menggunakan device yang akan mereka gunakan, dan dalam penelitian ini, penulis menggunakan laptop sebagai *device* utama, pada bagian awal dari sistem ini, user diharuskan membuka sistem ini, setelah itu user akan melihat bebarapa menu seperti *Home*, halaman dataset, dan halaman *predict*. *User* kemudian memilih halaman *predict.* Dalam halaman *predict user* diharuskan memasukkan judul yang akan di prediksi. Setelah *user* memasukkan judul, dengan menekan tombol prediksi selanjuntya adalah proses klasifikasi dimana sistem mengklasifikasi judul dengan metode *K-Nearest Neighbor.* Sistem ini dapat melakukan prediksi judul publikasi dari lima bidang ilmu yang disebutkan, maka alur proses dari sistem ini telah selesai.

3.2.2 Analisis Kebutuhan

Dalam tahapan Analisis kebutuhan adalah tahapan dimana sistem ini dianalisa tentang apa saja yang sistem ini lakukan dalam melakukan proses input sampai dengan mengeluarkan hasil dari sebuah klasifikasi bidang ilmu publikasi terindeks WoS yang dilakukan oleh sistem ini, dan ada beberapa proses atau fungsi yang harus ada pada sistem ini, diantara nya sebagai berikut:

A. Input Judul

Input judul adalah hal pertama yang harus dilakukan sistem ini nantinya, karena sistem ini membutuhkan sebuah judul yang dapat digunakan untuk melakukan klasifikasi. Sebelum melanjutkan proses langkah lainnya, untuk melakukan input judul, sistem ini sudah menyimpan dataset judul publikasi berupa format csv.

B. Proses klasifikasi

Fungsi kedua yang harus dimiliki oleh sistem ini adalah fungsi klasifikasi dimana sistem dapat mengolah dataset berupa judul publikasi sesuai dengan bidang ilmu menggunakan Metode KNN.

C. Melakukan prediksi

Dalam fungsi tersebut sistem bertugas untuk melakukan prediksi dari dataset judul publikasi yang telah dikumpulkan. Sistem ini akan melakukan pencarian tetangga terdekat terhadap judul publikasi yang akan diprediksi, dimana dalam mencari tetangga terdekat sistem ini akan menggunakan metode *K-Nearest Neighbor*.

D. Menampilkan hasil prediksi

Dan pada fungsi terakhir dari sistem ini adalah menampilkan hasil dari prediksi yang telah dilakukan sebelumnya, fungsi menampilkan hasil prediksi bertujuan untuk membuat user melihat hasil perhitungan yang telah diproses oleh sistem ini, pada fungsi ini user akan melihat berapa persentase tetangga terdekat dari setiap kelas yang dimiliki sesuai 5 bidang ilmu yang diprediksi.

3.2.3 Implementasi Sistem

Pada tahap implementasi sistem, akan dianalisa, beberapa *tools* yang digunakan kedepannya untuk pengembangan sistem ini, dan berikut merupakan *tools* yang digunakan untuk mengembangkan sistem ini:

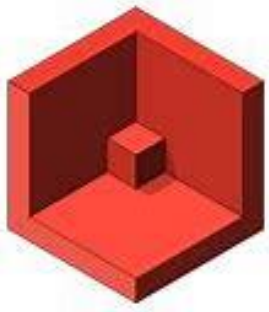
1. Perangkat Lunak

A. Visual Studio Code

Gambar 3. 27 Logo *Visual Studio Code*

Visual Studio Code atau VSC adalah sebuah media pengeditan yang dikembangkan *Microsoft.* Untuk menjalankan aplikasi tersebut, dapat menggunakan seperti *Windows, Linux*, dan *macOS*. Aplikasi ini memudahkan penulisan kode yang menunjang berbagai jenis pemrograman, meliputi *C++*, *Java*, *Python*, dan lainnya. Aplikasi dapat mendeteksi berbagai jenis bahasa pemrograman untuk memberikan variasi warna berdasarkan kemampuan pembuat kode. *Visual Studio Code* juga melekat dengan Github. Selanjutnya, dapat menambahkan plugin, di mana pengembang dapat menambahkan plugin untuk fungsionalitas yang tidak tersedia dalam aplikasi *Visual Studio Code*.

B. Rubixml



Gambar 3. 28 Logo *Rubixml Machine Learning.*

Rubix ML adalah *library PHP* yang dapat digunakan untuk membangun pembelajaran mendalam. *Library* ini menyediakan berbagai alat untuk membangun sistem ML, mulai dari pembacaan data, *preprocessing*, model training hingga testing. Semua fungsi tersebut dapat ditemukan di halaman *rubixml*.

C. Bootstrap

Bootstrap adalah *CSS framework* yang dikembangkan sebagai media membangun sebuah *front-end* dari web. Bootstrap merupakan framework *CSS* yang terkenal di kalangan programmer, terutama pengembang web atau (*web developer*).

D. PHP

Php merupakan bahasa pemrograman pendukung HTML yang berfungsi membuat aplikasi secara mudah yang memungkinkan memproses data dan pengolahan data. Php adalah perangkat lunak *Open Source*. PHP dikenal sebagai Bahasa *scripting*. Selanjutnya, PHP adalah bahasa *scripting* disimpan dan diproses oleh *server*. Semua sintaks yang diberikan akan diolah sepenuhnya deserver sementara hasilnya dikirim ke browser. Hasilnya dikirim ke pengguna, dimana pengguna yang menggunakan browser terintegrasi dengan tag HTML, yang diproses diserver dan dapat diperuntukkan untuk membuat halaman web yang mudah seperti *Active Server Pages* (ASP) atau *Java Server Pages* (JSP).

2. Perangkat Keras

Gambar 3. 29 Perangkat keras.

Dalam penelitian ini perangkat keras yang digunakan untuk menunjang penelitian yaitu Laptop VivoBook 14_Asus X441UB memiliki spesifikasi seperti processor Intel(R) dengan core i3-6006U, CPU 2.0GHz, 4GB RAM, Operasi Sistem Windows 10 Pro 64bit, dan sudah dilengkapi SSD 128G.

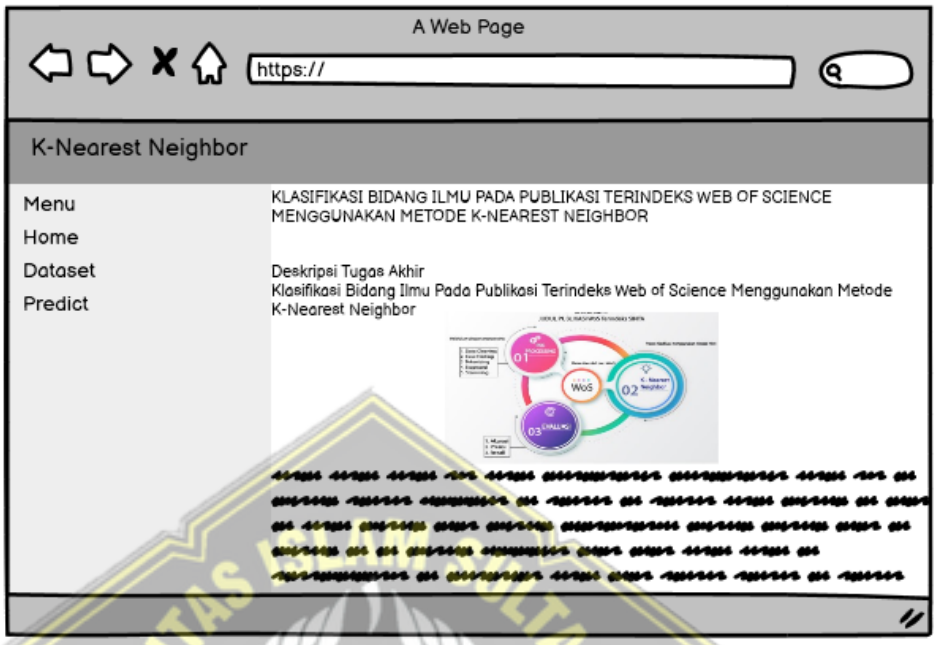
3.3 Perancangan Antarmuka

Pada bagian perancangan Antarmuka, pengguna merupakan mekanisme komunikasi antara pengguna dengan sistem. desain tampilan antarmuka berfungsi untuk mengetahui bentuk tampilan yang akan digunakan dalam aplikasi. Tampilan yang sederhana dan menarik, memberikan nilai tambah pada aplikasi. Berikut desain mockup yang akan dibuat dalam sistem adalah rancangan antarmuka dari Sistem klasifikasi bidang ilmu pada publikasi terindeks *Web of Science* menggunakan Metode K-*Nearest Neighbor*.

Gambar 3. 30 Tampilan sistem.

Pada gambar 3.30 yaitu tampilan sistem merupakan halaman berisikan menu yang ada pada sistem ini seperti Halaman Home, Halaman Dataset dan Halaman Predict.

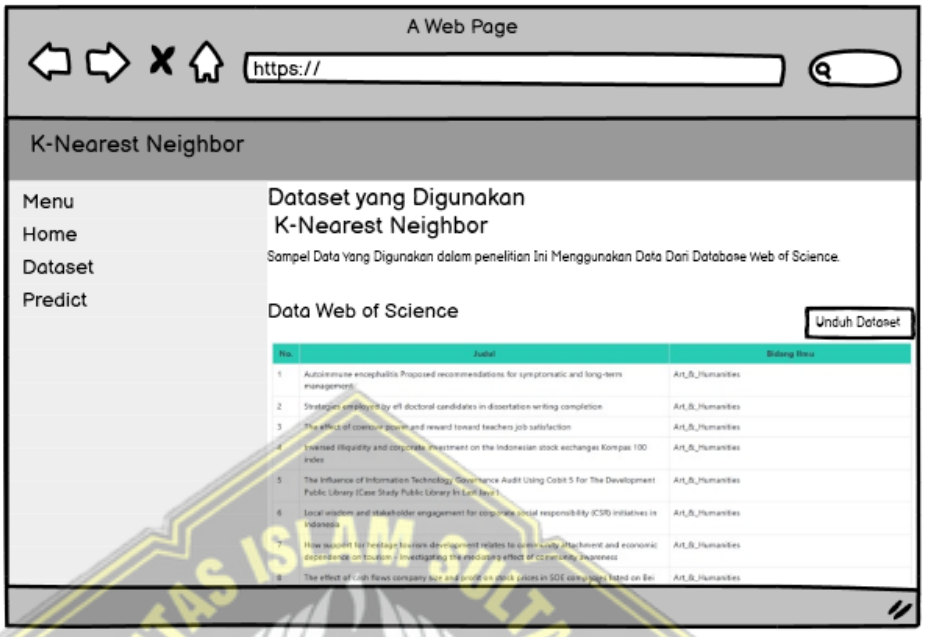
1. **Tampilan *Menu Home***



Gambar 3. 31 Tampilan *Menu Home*

Pada gambar 3.31 Halaman *Menu Home* merupakan halaman utama pada sistem ini. Halaman ini berisikan deskripsi tugas akhir penulis beserta tujuan pembuatan sistem ini. Penulis menggunakan metode *K-Nearest Neighbor*, di mana sistem dapat mengklasifikasikan publikasi terindeks WoS sesuai dengan lima bidang ilmu.

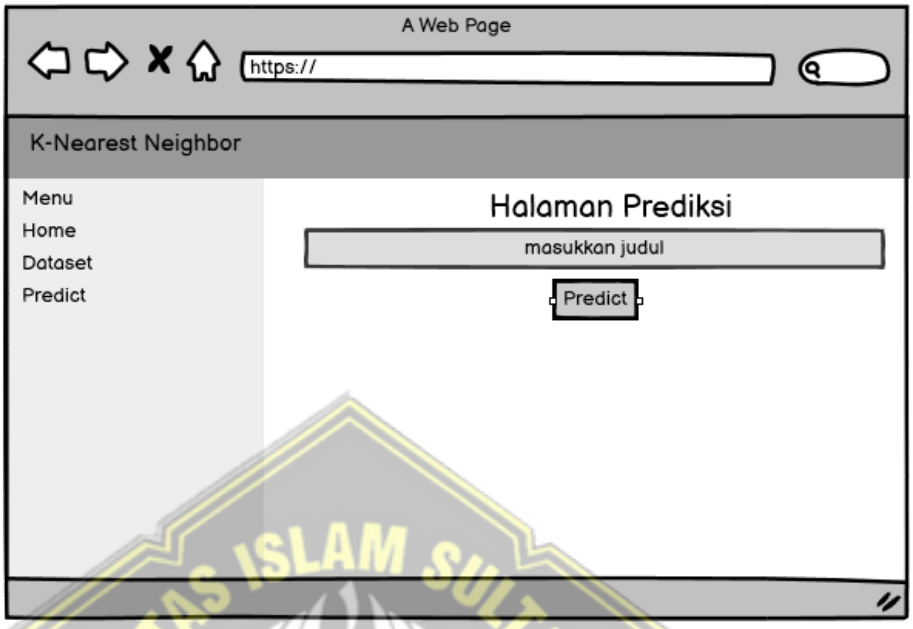
2. Tampilan *Menu* Dataset



Gambar 3. 32 Tampilan *Menu* Dataset.

Pada gambar 3.32 tampilan halaman ini, pengguna dapat melihat sampel kumpulan data-data yang digunakan penulis dalam penyusunan sistem. Penulis menggunakan data dari SINTA yang terindeks WoS. Halaman ini juga menawarkan unduh dataset yang berfungsi mengunduh file.

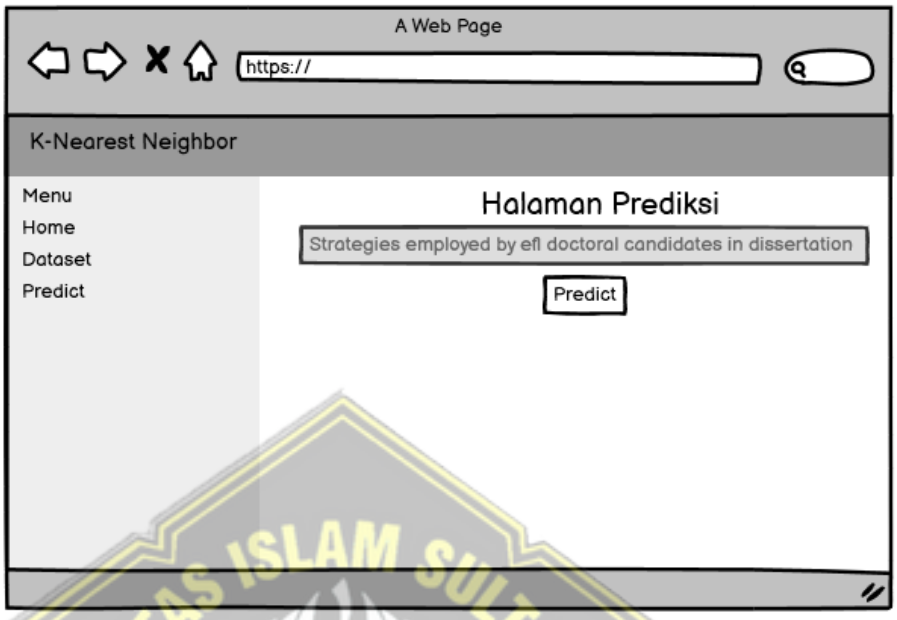
3. Tampilan *Menu Predict*



Gambar 3. 33 Tampilan *Menu Predict.*

Pada gambar 3.33 Halaman *predict* yaitu,bagian utama dari sistem ini, karena penulis berfokus membuat sebuah sistem yang sederhana dan mudah digunakan, maka halaman predict, dijadikan sebuah tempat yang dapat menjalankan 3 fungsi utama dari fungsi utama sistem ini yaitu, memasukkan judul,melakukan prediksi, serta menampilkan prediksi.

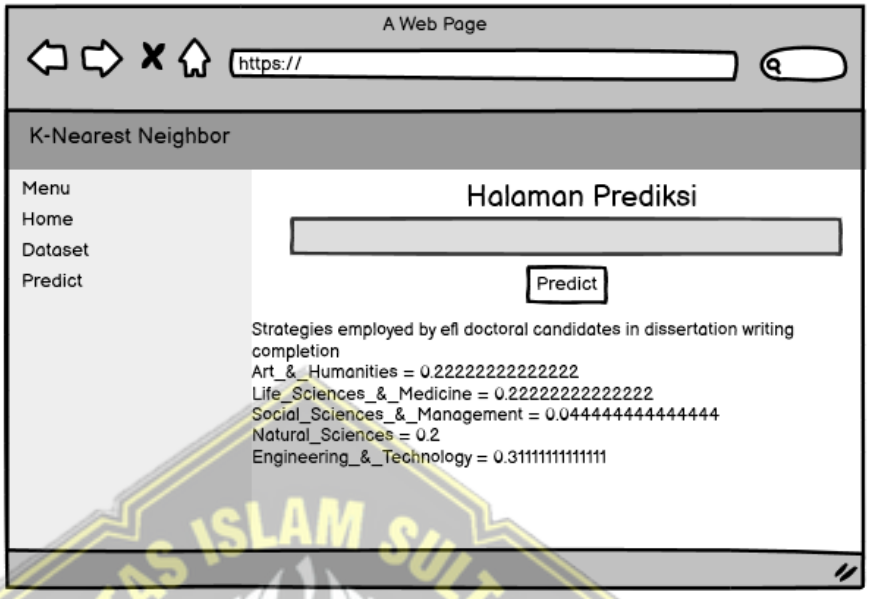
4. Tampilan *Menu Predict* saat memasukkan Judul



Gambar 3. 34 Tampilan memasukkan Judul.

Pada gambar 3.34 merupakan tampilan sebuah *menu predict* didalam halaman tersebut terdapat fungsi tombol memasukkan judul. *User* dapat menggunakan tombol tersebut untuk memasukkan judul yang diinginkan.

5. Tampilan *Menu Predict* pada Hasil Prediksi.



Gambar 3. 35 Tampilan Hasil Prediksi.

Pada gambar 3.35 adalah tampilan hasil dari tombol predict. Jadi, setelah user memasukkan judul lalu klik tombol predict. Maka output dari fungsi tersebut yaitu berupa hasil nilai akurasi masing-masing bidang.

BAB IV
HASIL DAN ANALISIS PENELITIAN

4.1 User Interface dan Penggunaan Sistem

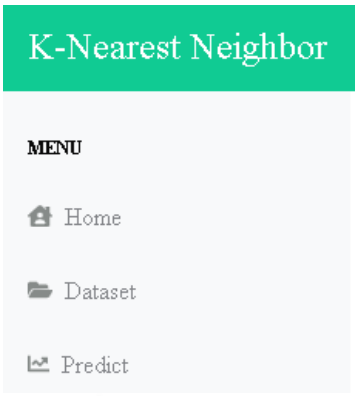
Tampilan kepada *user* atau bisa disebut *user interface*, adalah satu hal yang penting, karena pada bagian inilah yang akan sering dilihat dan berinteraksi, dan berikut adalah tampilan dari *user interface* dan Penggunaan Sistem dari sistem berbasis web yang telah dibuat oleh peneliti:

A. Tampilan *Menu Home*

Gambar 4. 1 Tampilan *Menu Home.*

Menu pertama adalah *Home*. Pada halaman ini menampilkan gambar dan tulisan penjelasan dari deksripsi tugas akhir, *user* hanya dapat melihat bagaimana tahapan sistem yang dibuat peneliti beserta penjelasan deskripsi tugas akhir dari sistem tersebut.

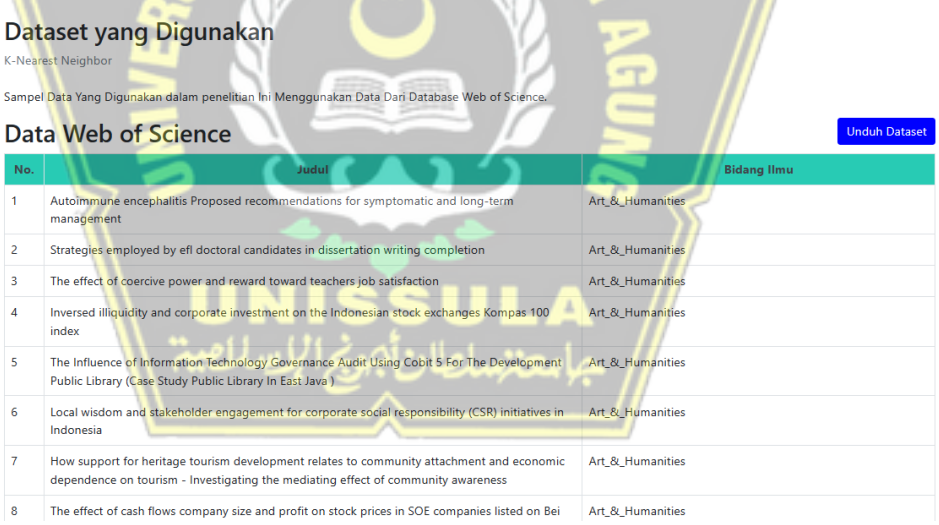
B. Tampilan *Menu*



Gambar 4. 2 Tampilan *Menu.*

Berikut menu merupakan pilihan fitur yang digunakan didalam sistem. Pada sistem terdapat berbagai fitur antara lain yaitu halaman *Home*, Halaman Dataset, dan Halaman *Predict*.

C. Tampilan *Menu* Dataset

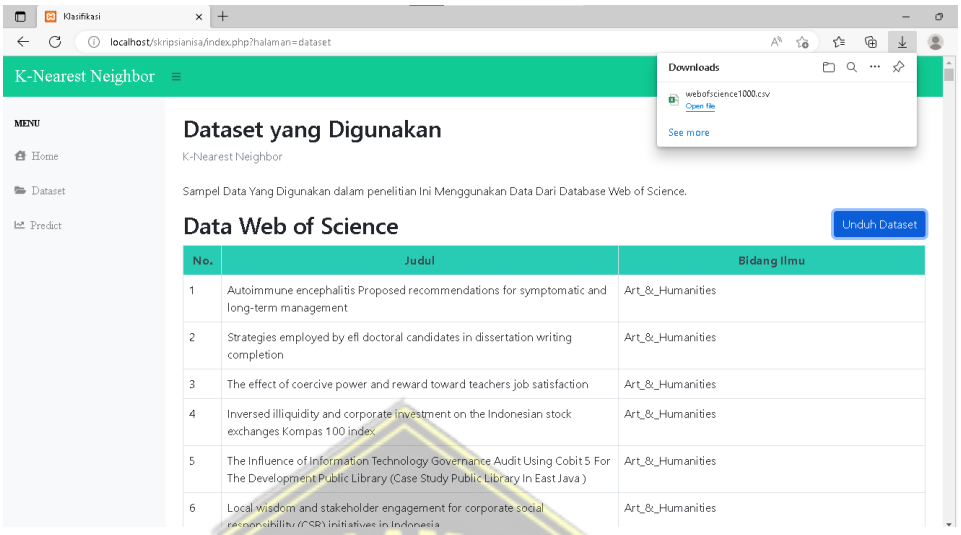


No.	Judul	Bidang Ilmu
1	Autoimmune encephalitis Proposed recommendations for symptomatic and long-term management	Art_&_Humanities
2	Strategies employed by efl doctoral candidates in dissertation writing completion	Art_&_Humanities
3	The effect of coercive power and reward toward teachers job satisfaction	Art_&_Humanities
4	Inversed illiquidity and corporate investment on the Indonesian stock exchanges Kompas 100 index	Art_&_Humanities
5	The Influence of Information Technology Governance Audit Using Cobit 5 For The Development Public Library (Case Study Public Library In East Java)	Art_&_Humanities
6	Local wisdom and stakeholder engagement for corporate social responsibility (CSR) initiatives in Indonesia	Art_&_Humanities
7	How support for heritage tourism development relates to community attachment and economic dependence on tourism - Investigating the mediating effect of community awareness	Art_&_Humanities
8	The effect of cash flows company size and profit on stock prices in SOE companies listed on Bei	Art_&_Humanities

Gambar 4. 3 Tampilan *Menu dataset.*

Menu kedua adalah Dataset. Pada halaman ini memiliki fitur unduh dan menampilkan tabel berisikan data *Web of Science* dimana tabel menjabarkan judul dari masing-masing 5 bidang ilmu. *User* dapat melihat hasil data pada tabel dan mengunduh dataset tersebut pada tombol yang telah disediakan. Hasil dari unduhan berupa format *excel*.

D. Tampilan Halaman Unduh Dataset

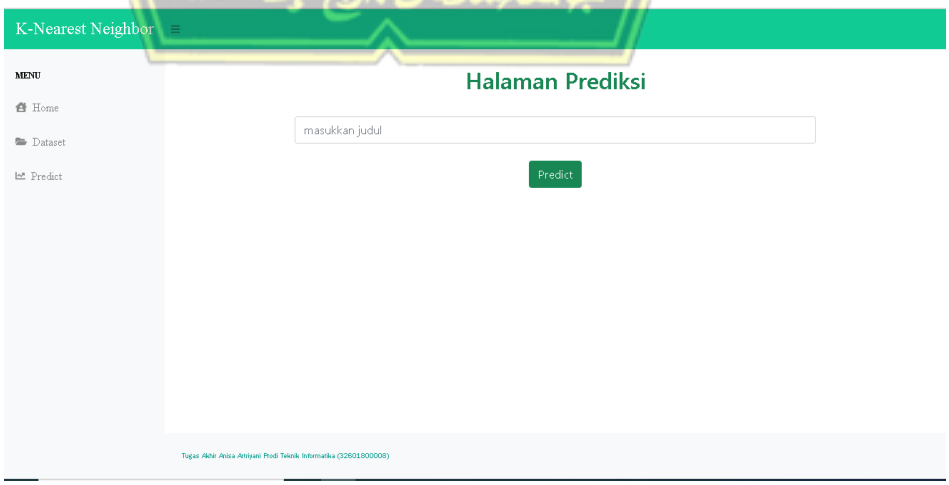


Gambar 4. 4 Tampilan Unduh *Dataset.*

Halaman Unduh Dataset adalah halaman dimana terdapat tombol unduh dataset yang mana tombol tersebut dapat mengunduh file dan melihat unduhan file dataset berupa format csv.

E. Tampilan *Menu Predict*

Menu terakhir adalah menu predict. Halaman ini memiliki fitur menampilkan hasil data predict yaitu dengan cara *user* memasukkan judul yang diinginkan kemudian klik tombol *predict* untuk melihat hasil dari data tersebut. Hasil data berupa judul, akurasi, dan bidang ilmu nya.



Gambar 4. 5 Tampilan *Menu predict.*

4.2 Analisa dan Pengujian

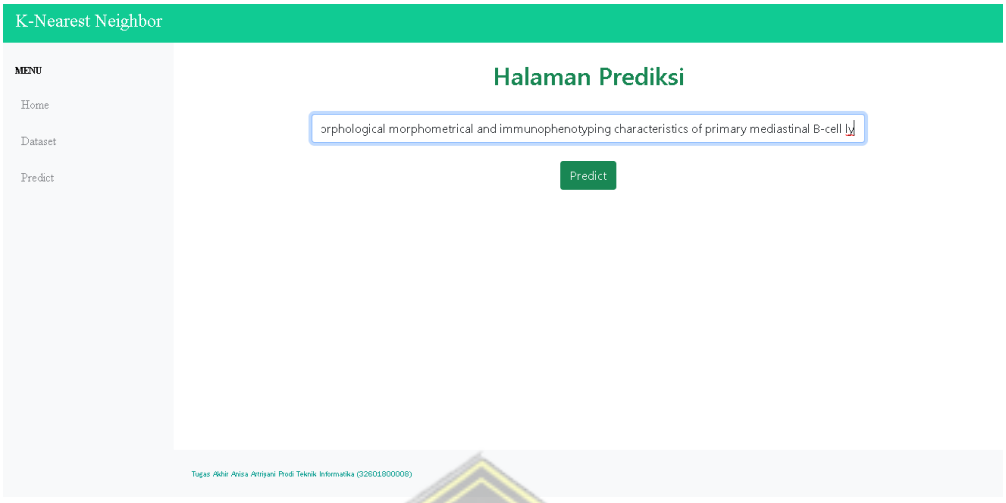
Pada tahapan ini dilakukan pengujian sistem dimana metode tersebut digunakan. Penelitian ini menggunakan pengujian yaitu *black box testing* dengan menggunakan *black box testing.* Pengujian tersebut adalah uji perangkat lunak dengan tujuan uji fitur aplikasi tanpa melihatnya dalam struktur atau cara kerja internal. Ada berbagai jenis kotak hitam tes seperti tes fungsional, tes non fungsional dan tes regresi, dan pengujian sistem menggunakan pengujian fungsional kali ini, di mana penelitian ini dalam pengujian setiap fitur yang tersedia pada sistem dan dalam tabel 4.1 hasil uji fungsional sistem ini.

Tabel 4. 1 Tabel pengujian.

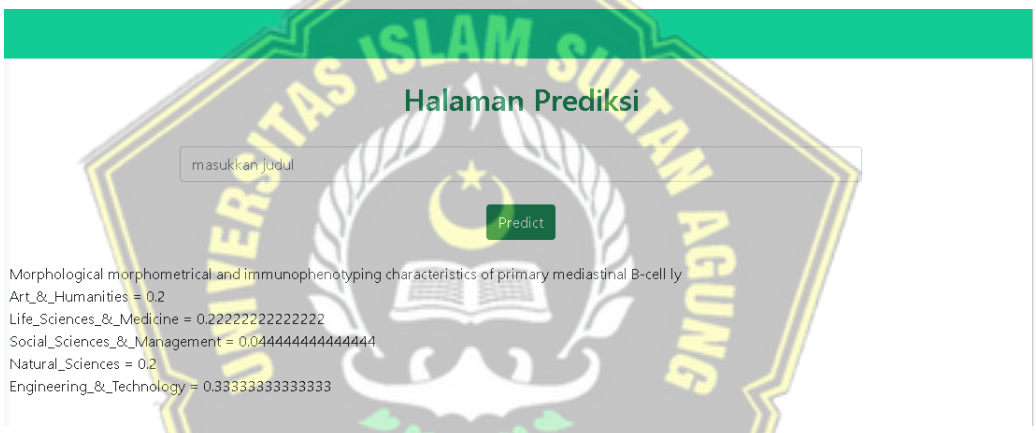
Skenario Pengujian	Kasus Pengujian	Hasil Pengujian	Kesimpulan
Unduh Dataset	Mengunduh Sampel Data.	Berhasil	Sesuai
Memasukkan Judul	Memasukkan judul publikasi.	Berhasil	Sesuai
Memprediksi Judul	Sistem memprediksi judul.	Berhasil	Sesuai
Hasil Prediksi Judul	*Output* hasil prediksi judul.	Berhasil	Sesuai

Dari hasil diatas dapat diketahui bahwa pengujian dilakukan sistem Klasifikasi bidang ilmu pada publikasi terindeks Wos menggunakan Metode KNN, sudah sesuai dan dapat dijalankan setiap fungsinya. Tahapan testing berjalan dengan baik sesuai dengan fungsi utama.

Selanjutnya, pada bagian *output* dari hasil prediksi klasifikasi ditemukan pengklasifikasian bidang ilmu, yang dimana menyebabkan penunjukan hasil akurasi tidak sesuai. Pada suatu ketika memasukkan judul dengan bidang ilmu tertentu. Hasil prediksi menunjukan nilai akurasi pada bidang ilmu tidak sesuai harapan. Sehingga nilai akurasi ini membingungkan pengguna, dimana output atau result yang dihasilkan akan menampilkan hasil sebagai berikut:



Gambar 4. 6 Tampilan memasukkan judul.



Gambar 4. 7 Tampilan Hasil Prediksi.

Output ini tentu tidak sesuai dengan hasil yang diharapkan, karena judul yang dimasukkan dengan bidang ilmu tertentu. melalui tahapan klasifikasi judul tersebut menunjukan hasil akurasi tertinggi pada bidang ilmu yang lain. Dengan adanya hambatan tersebut maka langkah yang di ambil selanjutnya yaitu memilah data sesuai dengan kelima bidang ilmu tersebut.

4.3 Analisa Akurasi

Tahap Analisa Akurasi bertujuan untuk dapat mengetahui seberapa tingkat akurasi sistem yang telah dibuat ini, tahap uji akan menggunakan data testing, yang mana berikut adalah pembagian dari data sampel dan data *testing*.

Tabel 4. 2 Pembagian data *testing* dan data *training*.

Nama Data	Jumlah	Jumlah Angka
Data Sampel	90%	900
Data Testing	10%	100

Total Data = 1000

Tabel 4. 3 Rincian Data

Class	*Art & Humanities*	*Engineering & Technology*	*Life Science & Medicine*	*Natural Sciences*	*Social Sciences & Management*
Data *Training*	200	200	200	200	200
Data *Testing*	20	20	20	20	20

Data tersebut kedepannya akan menghasilkan akurasi berupa *Confusion Matrix*. *Confusion matrix* dapat digunakan mencari nilai akurasi, presisi, dan *recall*. Hasil *training* yang di uji coba dapat menghasilkan *Confusion Matrix* berupa nilai akurasi, presisi, dan *recall.*

Dari perhitungan akurasi menghasilkan *Confusion Matrix* dimana *Confusion Matrix* ini digunakan sebagai pilihan mencari presisi, akurasi, dan recall. Variasi K yang berbeda yaitu 15,25,35,45,55 kemudian dicari nilai K paling mengenali atau mengklasifikasikan judul publikasi.

Tabel 4. 4 Hasil perhitungan *Confusion Matrix* dari K=15.

	Actually Positive	Actually Negatif
Predicted Positive	78 (TP)	222 (FP)
Actually Negatif	222 (FN)	312 (TN)

Tabel 4. 5 Hasil perhitungan *Confusion Matrix* dari K=25.

	Actually Positive	Actually Negatif
Predicted Positive	83 (TP)	217 (FP)
Actually Negatif	217 (FN)	332 (TN)

Tabel 4. 6 Hasil perhitungan *Confusion Matrix* dari K=35.

	Actually Positive	Actually Negatif
Predicted Positive	68 (TP)	232 (FP)
Actually Negatif	232 (FN)	272 (TN)

Tabel 4. 7 Hasil perhitungan *Confusion Matrix* dari K=45.

	Actually Positive	Actually Negatif
Predicted Positive	73 (TP)	227 (FP)
Actually Negatif	227 (FN)	292 (TN)

Tabel 4. 8 Hasil perhitungan *Confusion Matrix* dari K=55

	Actually Positive	Actually Negatif
Predicted Positive	75 (TP)	225 (FP)
Actually Negatif	225 (FN)	300 (TN)

Dengan data yang tertera pada tabel diatas dan dengan pengujian nilai K berbeda, diketahui bahwa nilai akurasi, presisi, dan *recall* dari setiap nilai K, berdasarkan rumus yang ada didapatkan hasil seperti tabel 4.8 dibawah ini.

Tabel 4. 9 Hasil pengukuran akurasi, *recall*, dan presisi.

Jumlah K	Akurasi	Recall	Presisi
15	0.4801	0.2599	0.2009
25	0.5079	0.2766	0.2149
35	0.4541	0.2266	0.2922
45	0.4664	0.2433	0.2179
55	0.4985	0.25	0.3180

Berdasarkan dari pengujian yang sudah dilaksanakan dapat disimpulkan, pada pengujian diatas akurasi berdasarkan nilai K yang telah ditentukan sebelumnya, sistem ini memiliki hasil prediksi pada judul dengan bidang ilmu yang di klasifikasi terdapat akurasi yang cukup rendah, terendah pada nilai K 35 yaitu bernilai 0.45, dengan recall dan presisi masing- masing bernilai 0.22, dan 0.29. Sedangkan nilai akurasi tertinggi, pada nilai K yang bernilai 25 dengan nilai akurasi, presisi, dan recall yaitu 0.50, 0.27, dan 0.21, skor tersebut merupakan skor tertinggi menurut penulis.

Tabel 4. 10 Perhitungan *confusion metrix* terbaik pada nilai k = 25

	Actually Positive	Actually negatif
Predicted Positive	83	217
Actually Negatif	217	332

Dari tabel 4.9 merupakan hasil perhitungan *confusion metrix* dengan nilai terbaik dari nilai k = 25. Dengan nilai *True Positive* yaitu 83, kemudian *False Negative* adalah 217, *False Negative* dengan nilai 217 dan *True Negative* yaitu bernilai 332.

4.4 Analisis Hasil Akurasi

Pada hasil akurasi yang di tunjukkan tabel 4.9 menunjukan hasil yang belum sesuai dengan yang diharapkan. Hasil tersebut dapat dikelompokkan sangat rendah dimana hasil akurasinya dengan nilai k 35 bernilai 0,45 dan nilai presisi, *recall* masing-masing yaitu 0,22 dan 0,29. Dari hasil itu, banyak faktor yang menjadi rendahnya hasil akurasi sebagai berikut:

1. Data yang digunakan kondisinya masih belum sesuai dengan bidang ilmu pada mestinya. Dalam kata lain setiap data nilai keakuratan sesuai bidang ilmu masih rendah.
2. Pada tema judul yang digunakan tidak sesuai dengan bidang ilmunya, maka dari itu bisa menyebabkan data tersebut rendah.
3. Data memiliki tingkat kemiripan baik perkataan maupun kalimat pada beberapa bidang ilmu yang berbeda.

Tabel 4. 11 Judul yang memiliki tema yang tidak sesuai bidang ilmunya.

No	Judul Artikel	Bidang Ilmu Yang Terlabel	Bidang Ilmu Yang Seharusnya
1.	The representation of colonial discourse in Indonesian secondary education history textbooks during and after the New Order (1975-2013).	*Art & Humanities*	*Engineering & Technology*
2.	Infrared-Assisted Extraction and HPLC-Analysis of Prunus armeniaca L. Pomace and Detoxified-Kernel a	*Natural Sciences*	*Engineering & Technology*

Pada tabel 4.11 Merupakan tabel yang berisi contoh judul yang dapat mempengaruhi nilai akurasi pada tabel tersebut berisi judul-judul yang memiliki tema tidak sesuai bidangnya salah satunya dengan judul bernama “The representation of colonial discourse in Indonesian secondary education history textbooks during and after the New Order (1975-2013)” yang mana judul tersebut telah terlabel *Art & Humanities* tetapi setelah di prediksi judul tersebut termasuk label *Engineering & Technology*. Tentu dengan penyebab rendahnya akurasi pada penelitian ini dan memungkinkan masih banyak faktor yang membuat rendahnya nilai akurasi seperti penggunaan bahasa yang mirip dengan kata dan kalimat satu lainnya.

4.5 Perbandingan Performa Data *Training* dan Data *Testing*


```
"accuracy": 0.490743277341215 5,
"balanced accuracy": 0.5082433356117566,
"f1 score": 0.15494393476044852,
"precision": 0.2118780096308186,
"recall": 0.24,
"specificity": 0.7764866712235132,
"negative predictive value": 0.5897727272727272,
"false discovery rate": 0.7881219903691814,
"miss rate": 0.76,
"fall out": 0.2235133287764867,
"false omission rate": 0.4102272727272728,
"mcc": 0.00697933502308544 2,
"informedness": 0.016486671223513306,
"markedness": -0.19834926309645412,
"true positives": 24,
"true negatives": 96,
"false positives": 76,
"false negatives": 76,
"cardinality": 100
```


Gambar 4. 8 *Output* Hasil dari data *Training.*

Pada gambar diatas yaitu 4.8 merupakan Output dari sampel data training dengan skenario 90%:10% yang berarti data training berjumlah 900 data. Dengan nilai akurasi, presisi, dan *recall* masing-masing berjumlah 0,49, 0,21, dan 0,24.

```
"accuracy": 0.43270083574961615,
"balanced accuracy": 0.48048245614035084,
"f1 score": 0.08205128205128205,
"precision": 0.07375886524822695,
"recall": 0.2,
"specificity": 0.7609649122807017,
"negative predictive value": 0.5001250781738587,
"false discovery rate": 0.926241134751773,
"miss rate": 0.8,
"fall out": 0.23903508771929824,
"false omission rate": 0.49987492182614135,
"mcc": -0.06689843447818355,
"informedness": -0.03903508771929824,
"markedness": -0.4261160565779144,
"true positives": 20,
"true negatives": 80,
"false positives": 80,
"false negatives": 80,
"cardinality": 100
```


Gambar 4. 9 *Output* hasil dari Data *Testing*.

Pada gambar 4.9 merupakan sebuah *Output* dari Data *Testing* dengan skenario 90:10 dengan jumlah data *testing* yaitu 100 data. Dimana nilai akurasi 0,43, nilai presisi adalah 0,07, dan nilai *recall* 0,2.

Tabel 4. 12 Perbandingan Data *Training* dan Data *Testing.*

	Akurasi	Presisi	*Recall*
Data Training	0,49	0,21	0,24
Data Testing	0,43	0,07	0,2

Dari tabel 4.11merupakan tabel yang berisi nilai masing-masing dari hasil *Training* dan *Testing.* Dengan data yang tertera pada tabel diatas dengan pengujian nilai K dan skenario uji yaitu 90:10 dapat diketahui nilai akurasi, presisi, dan *recall* dari nilai K yang ditentukan, berdasarkan rumus yang ada pada data *Training* mendapatkan hasil akurasi, presisi, dan *recall* bejumlah 0,49, 0,21, dan 0,24. Sedangkan hasil dari data *testing* memiliki jumlah nilai akurasi 0,43, nilai presisi adalah 0,07, dan nilai *recall* 0,2. Berdasarkan dari pengujian yang telah dilakukan dapat disimpulkan, bahwa pengujian akurasi berdasarkan nilai K dan dengan skenario 90%:10% yang telah ditentukan sebelumnya, yaitu Data *Training* dengan jumlah 900 dan Data *Testing* 100 data. Diketahui bahwa hasil dari Data tersebut, data Testing memiliki hasil yang rendah dibandingkan hasil Data *Training.* Hal ini menyebabkan *Overfitting*, *Overfitting* memiliki pengertian suatu keadaan dimana data yang digunakan untuk pengujian adalah data terbaik dari seluruh sampel data

tersedia. Sehingga, jika pengujian dilakukan dengan menggunakan data yang berbeda dapat mempengaruhi nilai akurasi dimana hasil yang diperoleh tidak sesuai harapan.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari penelitian ini dapat diambil kesimpulan bahwa klasifikasi bidang ilmu pada publikasi terindeks *Web of Science* menggunakan metode *K-Nearest Neighbor*. Hasil dari pengujian sistem dapat disimpulkan dengan akurasi, recall, dan presisi menghasilkan nilai terbaik pada nilai K 25 dengan nilai akurasi 0.50, *recall* yaitu 0.27, dan presisi sebesar 0.21. dan nilai terburuk pada nilai K 35 masing – masing nilai akurasi, presisi *recall* sebesar 0.45, 0,22, dan 0,29.

Dari hasil tersebut masih cukup rendah jika digunakan p ada sistem SINTA dikarenakan dataset yang di dapatkan kurang baik dan hasil klasifikasi tidak sesuai. Percobaan diatas dataset yang digunakan tidak sesuai harapan. Akan tetapi, sistem berjalan sebagaimana fungsinya, dan memberikan hasil nilai akurasi yang rendah sehingga perlu adanya peningkatan kualitas dari dataset.

5.2 Saran

Saran untuk perangkat sistem:

1. Untuk penelitian selanjutnya, perlu adanya tambahan *algoritma* lain untuk menunjang hasil keakuratan suatu hasil yang lebih baik.
2. Untuk pengembangan sistem berikutnya, penelitian disarankan menggunakan metode *K-Nearest Neighbor* dengan pencarian Nilai K secara *looping*.
3. Untuk penelitian selanjutnya perlunya memilah dataset yang akan diolah sehingga mampu menghasilkan data bersih.
4. Karena dari hasil akurasi keseluruhan dengan nilai K yang berbeda, dengan hasil yang masih rendah dan perlu tingkatkan pada penelitian berikutnya dapat digunakan algoritma selain *K-Nearest Neighbor*.
5. Pengembangan pada sistem bisa ditambahkan dengan berbagai fitur.

DAFTAR PUSTAKA

Nisha, A. C. (2021). Klasifikasi Abstrak Jurnal Repositor di Teknik Informatika UMM Menggunakan Metode Neighbor Weighted K-Nearest Neighbor. *Jurnal Repositor*, *3*(3), 295–304. https://doi.org/10.22219/repositor.v2i3.1225

Adhi Putra, A. D. (2021). Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *8*(2), 636–646. https://doi.org/10.35957/jatisi.v8i2.962

Maricar, M. A., & Dian Pramana. (2019). Perbandingan Akurasi Naïve Bayes dan K-Nearest Neighbor pada Klasifikasi untuk Meramalkan Status Pekerjaan Alumni ITB STIKOM Bali. *Jurnal Sistem Dan Informatika (JSI)*, *14*(1), 16–22. https://doi.org/10.30864/jsi.v14i1.233

Nabila, Z., Rahman Isnain, A., & Abidin, Z. (2021). Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(2), 100. http://jim.teknokrat.ac.id/index.php/JTSI

Hendrian, S. (2018). Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan. *Faktor Exacta*, *11*(3), 266–274. https://doi.org/10.30998/faktorexacta.v11i3.2777

Banyak, S., Untuk, S., Indonesia, B., Terbit, G. S., & Terbitan, P. (n.d.). *Jurus pertama : Memahami target publikasi Jurus kedua : Melakukan seleksi jurnal bersamaan dengan pelacakan referensi termutahir*. 1–11.

Fitria, A., & Azis, H. (2018). Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier. *Prosiding Seminar Nasional Ilmu Komputer Dan Teknologi Informasi*, *3*(2), 102–106.

Susanto, H., & Sudiyatno, S. (2014). Data mining untuk memprediksi prestasi siswa berdasarkan sosial ekonomi, motivasi, kedisiplinan dan prestasi masa lalu. *Jurnal Pendidikan Vokasi*, *4*(2), 222–231. https://doi.org/10.21831/jpv.v4i2.2547

Puspita, R., & Widodo, A. (2021). Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS. *Jurnal Informatika Universitas Pamulang*, *5*(4), 646. https://doi.org/10.32493/informatika.v5i4.7622

Utami, A. (2018). *ANALISIS INTEGRASI DATA SCOPUS PADA SINTA (Science and Technology Index) UNIVERSITAS SRIWIJAYA TAHUN 2018*.

Palma, B. K., Murdiansyah, D. T., & Astuti, W. (2021). Klasifikasi Teks Artikel Berita Hoaks Covid-19 dengan Menggunakan Algotrima K- Nearest Neighbor.https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15672.

Yuli Mardi. (2019). Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika. *Jurnal Edik Informatika*, *2*(2), 213–219.

Firmansyah, A., Qadri, R. A., & Arham, A. (2020). Pelatihan melalui Web Seminar terkait Publikasi Artikel untuk Menembus Jurnal Sinta 2 dan Scopus. *Abdimas: Jurnal Pengabdian Masyarakat Universitas Merdeka Malang*, *5*(2). https://doi.org/10.26905/abdimas.v5i2.4244

Kasanah, A. N., Muladi, M., & Pujianto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 196–201. https://doi.org/10.29207/resti.v3i2.945

SIADNYANI, L. (2018). Analisa Integrasi Data Sinta (Science and Technology Index) Menggunakan Website Internasional Dengan Manajamen Sistem Informasi Eis (Executive Information System) https://doi.org/10.1016/j.foeco.2018.06.029%

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *Jurnal Media Informatika Budidarma*, *5*(2), 640. https://doi.org/10.30865/mib.v5i2.2937

Devita, R. N., Herwanto, H. W., & Wibawa, A. P. (2018). Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(4), 427. https://doi.org/10.25126/jtiik.201854773

Sepdela, R. (2018). *Analisis Integrasi Data Dan Keterkaitan Website Sinta (Science and Technology Index) Ristekdikti Dengan Pangkalan Data Scopus*.

Ma'rifah, H., Wibawa, A. P., & Akbar, M. I. (2020). Klasifikasi Artikel Ilmiah Dengan Berbagai Skenario Preprocessing. *Sains, Aplikasi, Komputasi Dan Teknologi Informasi*, *2*(2), 70. https://doi.org/10.30872/jsakti.v2i2.2681

Rahmat Dian Nugraha, A., Auliasari, K., & Agus Pranoto, Y. (2020). IMPLEMENTASI METODE K-NEAREST NEIGHBOR (KNN) UNTUK SELEKSI CALON KARYAWAN BARU (Studi Kasus : BFI Finance

Surabaya). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *4*(2), 14–20. https://doi.org/10.36040/jati.v4i2.2656

Rozi, F. N., & Sulistyawati, D. H. (2019). Klasifikasi Berita Hoax Pilpres Menggunakan Metode Modified K-Nearest Neighbor Dan Pembobotan Menggunakan Tf-Idf. *Konvergensi*, *15*(1). https://doi.org/10.30996/konv.v15i1.2828