Kode/Nama Rumpun Ilmu : 458/Teknik Informatika

LAPORAN AKHIR PENELITIAN DOSEN

FEATURE SELECTION CHI-SQUARE DAN K-NN PADA PENGKATEGORIAN SOAL UJIAN BERDASARKAN COGNITIVE DOMAIN TAKSONOMI BLOOM

TIM PENGUSUL

1. **Indah Listiowarni., M.Kom** **(Ketua)**
   NIS : 710413580
2. **Eka Rahayu Setyaningsih.,M.Kom** **(Anggota)**
   NIS : 21023A347

UNIVERSITAS MADURA
2020

HALAMAN PENGESAHAN PENELITIAN DOSEN

Judul Penelitian	:	Feature Selection Chi-Square dan K-NN pada Pengkategorian Soal Ujian Berdasarkan Cognitive Domain Taksonomi Bloom
Kode/Nama Rumpun Ilmu	:	458 / Teknik Informatika
Ketua Peneliti	:	
a. Nama Lengkap	:	Indah Listiowarni,M.Kom
b. NIS	:	7104313580
c. Jabatan Fungsional	:	-
d. Program Studi	:	Teknik Informatika
e. Nomor HP	:	085730350676
f. Alamat surel	:	indah@unira.ac.id
Anggota Peneliti	:	
a. Nama Lengkap	:	Eka Rahayu Setyaningsih., M.Kom
b. Perguruan Tinggi	:	Sekolah Tinggi Teknik Surabaya
Biaya Penelitian	:	Rp. 3.000.000,-

Pamekasan, Mei 2018

Mengetahui,
Dekan Fakultas Teknik,

Ir.H.M.Hazin Mukti,MT, MM
NIP. 195906051987031002

Ketua Peneliti,

Indah Listiowarni, M.Kom
NIS. 7104313580

Menyetujui,
Ketua LPPM Universitas Madura

Dr. Zali, S.Pt.,M.Agr
NIP. 0706088401

DAFTAR ISI

halaman

Halaman Judul..... i
Lembar Pengesahan ii
Daftar Isi iii
Daftar Gambar iv
Daftar Tabel v
Abstrak..... vii
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Perumusan Masalah 2
1.3 Batasan Masalah 2
1.4 Tujuan Penelitian 3
1.5 Manfaat Penelitian..... 3
1.6 Sistematika Penulisan..... 3
BAB II LANDASAN TEORI..... 5
3.1. Data Mining 5
3.1. Klasifikasi 6
3.1. Tf-Idf 7
3.1. K-Nearest Neighbour 8
3.1. Chi-Square 9
3.1. Taksonomi Bloom 10
3.1. Tool Penunjang Perancangan Sistem..... 11
2..1. Flowchart..... 11
2..2. DFD..... 12
2..3. CDM..... 13
2..4. PDM..... 14
2.7 Apache HTTP Server..... 15
2.8 MySQL Database Server..... 15
2.9 MySQL 15

BAB III METODOLOGI 23
3.1. Dataset 23
3.2 Daftar Class 23
3.3 Alur Sistem 25
3.4 Tf-Idf 26
3.5 Chi-Square 26
3.6 K-Nearest Neighbour 27
3.7 Pre-Processing 28
3.8 Evaluasi 28
BAB IV HASIL DAN PEMBAHASAN 30
BAB V PENUTUP 35
5.1. Kesimpulan..... 35
5.2. Saran 35
DAFTAR PUSTAKA..... 36

DAFTAR GAMBAR

halaman

Gambar 2.1 Piramida Taksonomi Bloom Ranah Kognitif 11
Gambar 2.2 Contoh CDM 14..... 14
Gambar 3.1 Piramida level cognitive domain..... 24
Gambar 3.2. Alur sistem 25
Gambar.3.3 tahapan pre-processing 28
Gambar 4.1 Output Sistem 30
Gambar 4.2 pengaruh nilai K terhadap F-measure 33

DAFTAR TABEL

halaman

Tabel 2.1 Kotingensi Term dan Class..... 9
Tabel 2.2 tabel simbol flowchart 12
Tabel 2.3 tabel simbol DFD 13
Tabel 2.4 Tabel Simbol CDM 14
Tabel 3.1 daftar class dan contoh kata kerja operasional 24
Tabel 3.2. Tabel Kontingensi Term dan Class 27
Tabel 3.3. Tabel confusion matrix 29
Tabel 4.1. Data soal 30
Tabel 4.2.Perhitungan Korelasi 31
Tabel 4.3. Hasil Chi-square 31
Tabel 4.4. Hasil akhir Chi-square 32
Tabel 4.5. Nilai K pada KNN-chi 32
Tabel 4.6. Kombinasi data training 33
Tabel 4.7. Hasil ujicoba dataset pertama 34

ABSTRAK

Examination merupakan tolak ukur akhir bagi peserta didik dengan tujuan untuk mengidentifikasi kemampuan cognitive dan daya tangkap siswa selama menjalankan kegiatan belajar mengajar di kelas, sehingga ketika diadakan sebuah ujian diperlukan susunan soal yang tepat. Proses klasifikasi data soal secara manual sangat panjang dan rumit, karena juga berkaitan dengan banyaknya data dan pengecekan setiap term, sehingga dibutuhkan sebuah sistem yang secara otomatis dapat diterapkan untuk membuat proses klasifikasi data soal lebih mudah dan tidak menghabiskan banyak waktu. Taksonomi bloom merupakan sebuah taksonomi yang digunakan di dunia pendidikan berisi enam level cognitive berdasarkan tingkat kesulitannya. Penelitian ini berkonsentrasi pada pengkategorian data soal ujian biologi tingkat SMA berdasarkan cognitive domain taksonomi bloom, dan metode yang digunakan adalah K-Nearest Neighbour (KNN),dan metode feature selection Chi-Square (X2)yang bertujuan untuk menyeleksi fitur yang diperlukan. Dataset yang digunakan pada penelitian ini terdiri dari 2 macam dataset, yaitu pada dataset pertama metode yang digunakan berhasil mencapai nilai F-measure tertinggi 79,36%, sedangkan pada dataset kedua,berhasil mencapai nilai F-measure tertinggi 61,56%

Kata Kunci : KNN, taksonomi bloom, Chi-Square, F-measure

BAB I
PENDAHULUAN

2.1 Latar Belakang

Taksonomi bloom merupakan salah satu taksonomi yang digunakan dalam ranah pendidikan dan digunakan sebagai patokan untuk menyusun soal-soal yang digunakan pada ujian-ujian sekolah, misalnya Ulangan Harian, Ujian Tengah Semester,dan Ujian Akhir bahkan pada buku latihan soal lainnya. Taksonomi Bloom memiliki 3 domain yaitu afektif, cognitive dan psikomotorik. Pada penelitian ini akan berkonsentrasi pada domain cognitive yang berisi 6 level.Soal-soal ujian yang digunakan di dalam dunia pendidikan disesuaikan dan dikategorikan oleh 6 level tersebut, kemudian setiap jenis Ujian memiliki presentase tersendiri dalam peyusunan soal yang terdiri dari soal-soal yang telah terkategori ke dalam 6 level cognitive domain taksonomi bloom.

Kata “Bloom” pada Taksonomi Bloom merupakan nama ilmuwan yang pertama kali menggagas taksonomi ini, yaitu Benjamin Samuel Bloom pada tahun 1956, dan dilakukan revisi ulang oleh Lorin Anderson Krathwohl pada tahun 1994, sehingga terjadi sedikit perubahan pada keenam level sebelumnya. Pada penelitian ini akan digunakan 6 level cognitive domain taksonomi bloom yang telah direvisi untuk mengkategorikan data berupa soal-soal ujian biologi tingkat SMA. Enam level yang akan digunakan terdiri dari : Mengingat (remembering), Memahami (understanding),Menerapkan (applying),Menganalisis (analysing), Mengevaluasi (evaluating),Membuat (creating).

Pengkategorian soal ujian yang dilakukan secara manual, akan menghabiskan banyak waktu dan tenaga, dikarenakan rumitnya proses klasifikasi dan banyaknya data. Biasanya, Setiap soal yang akan dikategori dan dicari classnya berdasarkan cognitive domain taksonomi bloom akan dilihat setiap kata yang membangun soal tersebut, dibaca baik-baik, mengaitkan dengan kata kerja operasional yang tersedia, kemudian ditelaah maksud dari soal tersebut, dan akhirnya dikategorikan sesuai dengan enam level cognitive domain taksonomi bloom.

Sehingga, untuk memudahkan proses kategorisasi dibutuhkan sebuah sistem yang mampu mengkategorikan soal ujian secara otomatis. Membangun sebuah sistem, tidak luput dengan penggunaan metode. K-Nearest Neighbour merupakan algoritmayang dipilih dan akan digunakan pada penelitian ini, metode feature selection yang akan digunakan sebagai metode seleksi fitur adalah Chi-Square.Metode feature selection berguna untuk meneyeleksi fitur yang dianggap penting dan membuang fitur-fitur yang tidak diperlukan. Pada penelitian sebelumnya, K-Nearest Neighbour memiliki perfomance terbaik ketika digunakan bersama metode feature selection, dibandingkan dengan dua classifier pembandingnya, yaitu Naive Bayessian (NB) dan Support Vector Machine (SVM).

2.2 Perumusan Masalah

Berdasarkan latar belakang diatas, maka dapat dirumuskan permasalahan sebagai berikut :

1. Bagaimana Mengimplementasikan metode Feature Selection terhadap metode Klasifikasi untuk mengklasifikasikan data soal ke dalam 6 class taksonomi bloom
2. Bagaimana mengetahui Perfomance dan ketepatan klasifikasi menggunakan metode K-Nearest Neighbour dan Feature Selection Chi-Square.

2.3 Batasan Masalah

Adapun batasan masalah dalam tugas akhir ini, yaitu :

1. Data yang digunakan berjumlah 600 data, dengan menggunakan variasi data testing sebagai berikut, yaitu 10,25,50 dan 100 data
2. Metode yang digunakan adalah K-Nearest Neighbour sebagai metode klasifikasi dan Chi-Square sebagai metode Feature Selection
3. Untuk mengukur Perfomance digunakan metode F-Measure
4. Data yang digunakan adalah data text soal tingkat SMA

2.4 Tujuan Penelitian

Penulisan Tugas Akhir ini mempunyai beberapa tujuan antara lain adalah :

1. Membuat aplikasi Klasifikasi Soal biologi tingkat SMA ke ranah kognitif Taksonomi Bloom
2. Mengetahui Perfomance kombinasi metode KNN dan Chi-Square
3. Bagi penulis, untuk memenuhi kewajiban tridharma yaitu penelitian

2.5 Manfaat Penelitian

Adapun manfaat yang diperoleh dari hasil penelitian ini, diantaranya :

1. Memberikan kemudahan untuk meng-klasifikasikan data soal biologi ke dalam 6 class ranah kognitif taksonomi bloom
2. Memberi kemudahan dalam pengolahan data soal

1.6 Sistematika Penulisan

Sistematika penulisan ini dimaksudkan agar mempermudah proses pembahasan dalam mempelajari dan memahami isi pada bagian-bagian apa saja yang termuat dari bab dan sub bab. Berikut adalah sistematika penyusunan yang akan digunakan untuk mengembangkan tugas akhir ini:

Bab I Pendahuluan

Pada bab ini, menjelaskan dan menguraikan secara singkat mengenai Latar Belakang, Perumusan Masalah, Batasan Masalah, Tujuan Penelitian, Manfaat Penelitian, Metode Penelitian, Tinjauan Pustaka, Sistematika Penulisan Laporan Penelitian Dosen.

Bab II Landasan Teori

Bab ini menjelaskan tentang teori-teori yang mendukung permasalahan mengenai pengertian Metode Taksonomi Blooo, Klasifikasi dan metode-metode yang digunakan yaitu TF-idf, KNN, Chi-Square dan F-Measure, dan beberapa tool penunjan sistem seperti Flowchart, DFD dan CDM-PDM

Bab III Metodologi

Dalam bab ini dijelaskan mengenai metodologi meliputi beberapa hal, diantaranya pengumpulan data, analisis data, kebutuhan sistem serta perancangan sistem.

Bab IV Implementasi dan Pengujian Sistem

Bab ini berisi hasil dan pembahasan tentang pengujian aplikasi terhadap implementasi metode KNN dan chi-Square, visualiasi data dari hasil pengujian data, dan hasil performance.

Bab V Penutup

Kesimpulan merupakan jawaban atas pertanyaan-pertanyaan pada rumusan masalah dan intisari dari hasil penelitian. Sedangkan saran merupakan kumpulan dan rekomendasi dari penulis untuk mengembangkan sistem yang telah dibuat.

BAB II
LANDASAN TEORI

2.1 Data Mining

Data mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dan suatu basis data dengan melakukan pola-pola dari data dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga yang diperoleh dengan cara mengektraksi dan menggali pola yang penting atau menarik dari data yang terdapat dalam basis data. Data mining biasa juga dikenal nama lain seperti knowledge discovery in database (KDD) ekstraksi pengetahuan (knowledge extraction) analisa data/pola dan kecerdasan bisnis (business intelligence) dan merupakan alat untuk memenipulasi data untuk penyajian informasi sesuai kebutuhan pengguna dengan tujuan untuk membantu dalam analisis koleksi pengamatan perilaku.

Menurut Pramudio (2006) ,Data Mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Sedangkan menurut Turban (2005) data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar

Karakteristik data mining sebagai berikut :

1. Data mining berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
2. Data mining biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih dipercaya.
3. Data mining berguna untuk membuat keputusan yang kritis, terutama dalam strategi.

Berdasarkan beberapa pengertian tersebut dapat ditarik kesimpulan bahwa data mining adalah suatu teknik menggali informasi berharga yang terpendam atau tersembunyi pada suatu koleksi data (database) yang sangat besar sehingga ditemukan suatu pola yang menarik yang sebelumnya tidak diketahui.

Kata mining sendiri berarti usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasar. Karena itu data mining sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (artificial intelligent), machine learning, statistik dan database. Beberapa metode yang sering disebut-sebut dalam literatur data mining antara lain clustering, classification, association rules mining, neural network, genetic algorithm dan lain-lain

Menurut Larose (2005), terdapat enam fungsi dalam data mining, yaitu :

1. Fungsi Deskripsi (descrition)
2. Fungsi Estimasi (estimation)
3. Fungsi Prediksi (Prediction)
4. Fungsi Klasifikasi (classification)
5. Fungsi Pengelompokan (classification), dan
6. Fungsi Asosiasi (association)

Mengacu kepada Berry dan Browne (2006), keenam fungsi data mining tersebut dapat dipilah menjadi :

1. Fungsi minor atau fungsi tambahan, yang meliputi ketiga fungsi yang pertama, yaitu deskripsi, estimasi dan prediksi
2. Fungsi mayor atau fungsi utama , yang meliputi ketiga fungsi berikutnya, yaitu klasifikasi, pengelompokan, dan asosiasi

2.2 Klasifikasi

Klasifikasi adalah menentukan sebuah record data baru ke salah satu dari beberapa kategori (klas) yang telah didefinisikan sebelumnya (Fajar Astuti,2013). Klasifikasi adalah salah satu tugas yang penting dalam data mining, dalam klasifikasi, sebuah pengklasifikasi dibuat dari sekumpulan data latih dengan kelas yang telah ditentukan sebelumnya.

Klasifikasi data adalah proses dua langkah. Pada langkah pertama, sebuah model dibangun menggambarkan sebuah kumpulan kelas data atau konsep dari populasi data yang telah ditentukan sebelumnya (misalkan data pengajuan pinjaman bank). Model tersebut dibangun dengan menganalisa data latih yang digambarkan oleh atribut-atribut. Tiap tuple di asumsikan untuk dimiliki oleh kelas yang telah di tentukan, seperti di tentukan oleh salah satu atribut, yang dinamakan

class label attribute. Langkah kedua adalah menguji model yang telah dibangun kepada data uji

Untuk mengukur ketepatan atau performa model dalam mengklasifikasi data uji. Setelah pengukuran performa selesai dilakukan, pengambil keputusan dapat memutuskan untuk menggunakan model tersebut atau mengulang pembuatan model dengan data latih atau metode yang berbeda untuk menghasilkan model klasifikasi yang lebih baik.

Contoh yang bisa digunakan dalam klasifikasi data adalah data pengajuan pinjaman bank, dengan atribut-atribut nama, umur, pendapatan dan rating kredit dengan nama, umur, pendapatan sebagai data tuple dan seterusnya akan disebut sebagai atribut saja, sedangkan atribut rating kredit sebagai class label atribute dimana seterusnya akan disebut sebagai kelas. Nilai-nilai yang terdapat dalam data (seperti nama = Indah Listiowarni, umur = >40, pendapatan = Tinggi, dan rating_kredit = baik) disebut sebagai kumpulan nilai atribut, dimana seterusnya akan disebut sebagai kumpulan atribut saja. Aturan Klasifikasi yang dibuat dari kumpulan-kumpulan atribut pada data pelatihan dan Algoritma Klasifikasi disebut sebagai model klasifikasi.

Pendekatan umum yang digunakan dalam masalah klasifikasi adalah, pertama, training set berisi record yang mempunyai label kelas yang diketahui haruslah tersedia. Training set digunakan untuk membangun model klasifikasi, yang kemudian diaplikasikan ke test set, yang berisi record-record dengan label kelas yang tidak diketahui.

2.3 Tf-Idf

TF-IDF (Term Frequency-Inverse Document Frequency) merupakan metode yang biasa dipakai dalam implementasi text mining, salah satunya adalah digunakan untuk pembobotan kata sebagai strategi untuk mengkategorikan dokumen

Proses pembobotan kata menggunakan metode TF-IDF, dilakukan dengan menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF) pada setiap token (kata) di setiap dokumen dalam korpus. Metode ini akan menghitung bobot setiap token t di dokumen d dengan rumus:

$$w_{t,d} = tf_{t,d} \times idf_t \quad (1)$$

di mana tf adalah jumlah kemunculan setiap term t dalam sebuah dokumen d sedangkan idf merupakan kemunculan term t pada keseluruhan dokumen, atau juga bisa disebut pembobotan global, nilai idf didapatkan dengan menggunakan formula sebagai berikut:

$$idf_t = \log \frac{N}{df_t} \quad (2)$$

dft merupakan merupakan jumlah dokumen yang mengandung term t, N adalah jumlah dokumen dari keseluruhan data yang digunakan.

2.4 K-Nearest Neighbour

K-NN (K-nearest neighbour) merupakan classifier yang mengklasifikasikan data berdasarkan perbandingan K tetangga terdekat, parameter K pada KNN memiliki pengaruh dalam penetuan hasil prediksi, dimana parameter K tersebut tidak boleh melebihi jumlah data testing dan harus bilangan ganjil untuk menghindari data double. KNN termasuk supervised learning karena metode ini mengklasifikasikan data atau sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumnya (data training)

KNN mengukur similarity data menggunakan pengukuran distance, seperti Euclidian, Cosine similarity, Manhattan, Square Euclidian ,dan lain-lain. Pada penelitian ini digunakan Euclidean distance untuk menghitung similarity dari dua data atau lebih.

$$d_{(x_1,x_2)} = \sqrt{(|x_{11} - x_{21}|)^2 + (|x_{12} - x_{22}|)^2} \quad (3)$$

variabel d merupakan distance dari dua data, yaitu data x1 dan data x2. Selain dikenal sebagai metode yang sederhana dan mudah diimplementasikan, KNN juga dikenal sebagai metode yang membutuhkan waktu dan cost komputasi yang besar, karena harus menyocokkan data tes, dengan seluruh data training yang ada. Sehingga dibutuhkan adanya proses feature selection.

Secara garis besar, alur proses dari metode classifier K-Nearest Neighbour, meliputi langkah-langkah sebagai berikut :

1. Tentukan data testing

2. Tentukan jumlah tetangga terdekat K
3. Hitung jarak data testing ke data training
4. Urut data berdasarkan data yang mempunyai jarak terdekat
5. Tentukan kelompok data testing berdasarkan label mayoritas pada nilai k yang sudah ditentukan

KNN memiliki beberapa kelebihan yaitu ketangguhan terhadap training data yang memiliki banyak noise dan efektif apabila training data-nya besar. Sedangkan, kelemahan KNN adalah sulitnya menentukan jarak dan atribut yang harus digunakan untuk mendapatkan hasil terbaik, selain itu biaya komputasi yang digunaan cukup tinggi karena diperlukan perhitungan jarak setiap term data testing pada keseluruhan data training.

2.5 Chi-Square

Chi-Square atau yang biasa disimbolkan sebagai X2 merupakan metode feature selection atau pemilihan fitur yang termasuk dalam pendekatan metode filter. Metode ini menggunakan perhitungan statistik untuk mendapatkan term unik pada sebuah dokumen dan membuang term yang tidak relevant, sehingga mempersingkat waktu komputasi pada tahapan selanjutnya.

Penggunaan metode Chi-Square untuk pemilihan fitur pada dokumen, yaitu didasarkan pada dua variabel yang saling terkait yaitu term t dan class c. Perhitungan untuk mengetahui term yang berkontribusi pada class tertentu dapat menggunakan tabel kontingesi untuk memudahkan perhitungan, sebagai berikut pada tabel 2.1, dengan menggunakan tabel kontingensi dapat diketahui apakah term pada sebuah class tertentu dinyatakan independen atau dependen. Jika term dinyatakan dependen maka term tersebut menyerupai atau sama dengan label kategori.

Tabel 2.1

Tabel Kotingensi Term dan Class

	t	not t
Class	A	C
not Class	B	D

Sehingga dengan menggunakan tabel kontingensi pada tabel 2.1, didapatkan formula untuk perhitungan nilai Chi-Square

$$X^2(t,c) = \frac{N(A\times D - B\times C)^2}{(A+B)\times(C+D)\times(A+C)\times(B+D)} \quad (4)$$

variabel t merupakan kata yang sedang diujikan terhadap suatu kelas c, N merupakan jumlah dokumen latih, A merupakan banyaknya dokumen pada kelas c yang memuat kata t, B merupakan banyaknya dokumen yang tidak berada di c namun memuat kata t, C merupakan banyaknya dokumen yang berada di kelas c namun tidak memiliki kata t di dalamnya, serta D merupakan banyaknya dokumen yang bukan merupakan dokumen kelas c dan tidak memuat kata t

2.6 Taksonomi Bloom

Taksonomi berasal dari dua kata dalam bahasa Yunani yaitu tassein yang berarti mengklasifikasi dan nomos yang berarti aturan, jadi Taksonomi merupakan hierarki klasifikasi yang dijadikan sebuah dasar atau panutan. Taksonomi bloom merupakan sebuah aturan klasifikasi/kategori yang digagas oleh ilmuwan yang bernama Benjamin Samuel Bloom yang terdiri level-level dari tingkat yang rendah hingga ke levelyang tertinggi (hierarki). Bloom membagi kerangka konsep ini dalam tiga garis besar yaitu : kognitif, afektif dan psikomotorik.Ranah Kognitif berisi perilaku yang menekankan aspek intelektual, seperti pengetahuan, dan keterampilan berpikir. Ranah afektif mencakup perilaku terkait dengan emosi, misalnya perasaan, nilai, minat, motivasi, dan sikap. Sedangkan ranah Psikomotorik berisi perilaku yang menekankan fungsi manipulatif dan keterampilan motorik / kemampuan fisik, berenang, dan mengoperasikan mesin, jadi secara singkat Kognitif domain menekan kan pada knowledge, Afektif pada attitude dan Psikomotorik pada skills.

Ranah kognitif (cognitive domain) mengurutkan keahlian berpikir sesuai dengan tujuan yang diharapkan. Proses berpikir menggambarkan tahap berpikir yang harus dikuasai oleh siswa agar mampu mengaplikasikan teori kedalam perbuatan. Ranah kognitif ini terdiri atas enam level, yaitu: (1) knowledge (pengetahuan), (2) comprehension (pemahaman atau persepsi), (3) application

(penerapan), (4) analysis (penguraian atau penjabaran), (5) synthesis (pemaduan), dan (6) evaluation (penilaian).Ke-enam level pada ranah kognitif dapat digambarkan dengan menggunakan piramida, pada gambar 2.1

Gambar 2.1
Piramida Taksonomi Bloom Ranah Kognitif

2.7 *Tool* Penunjang Perancangan Sistem

Menurut Whitten, perancangan sistem adalah "Proses dimana keperluan pengguna dirubah ke dalam bentuk paket perangkat lunak dan atau kedalam spesifikasi pada komputer yang berdasarkan pada sistem informasi."

Tool Penunjang Perancangan Sistem adalah alat atau proses yang digunakan untuk menggambarkan proses perancangan sistem dalam bentuk diagram. Ada beberapa Tool Penunjang Perancangan Sistem yaitu sebagai berikut:

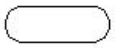
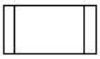
2.7.1 Flowchart

Flowchart merupakan gambar atau bagan yang memperlihatkan urutan dan hubungan antar proses beserta instruksinya. Gambaran ini dinyatakan dengan simbol. Dengan demikian setiap simbol menggambarkan proses tertentu. Sedangkan hubungan antar proses digambarkan dengan garis penghubung.

Flowchart ini merupakan langkah awal pembuatan program. Dengan adanya flowchart urutan poses kegiatan menjadi lebih jelas. Jika ada penambahan proses maka dapat dilakukan lebih mudah. Setelah flowchart selesai disusun, selanjutnya pemrogram (programmer) menerjemahkannya ke bentuk program dengan bahsa pemrograman.

Berikut adalah beberapa simbol yang digunakan dalam menggambar suatu flowchart dijelaskan pada table 2.2

Tabel 2.2 tabel simbol flowchart

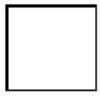
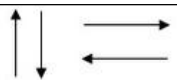
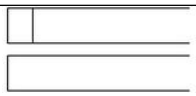
SIMBOL	NAMA	FUNGSI
	Terminator	Permulaan/akhir progran
	Garis Alir	Arah Aliran program
	Preparation	Proses inisialisasi harga awal
	Proses	Proses perhitungan/proses pengolahan data
	Input/output data	Proses input/output data,parameter,informasi
	Predefined process (sub program)	Permulaan sub program/proses menjalankan sub program
	Decision	Perbandingan pernyataan yang memberikan pilihan untuk langkah selanjutnya
	On page connector	Penghubung bagian flowchart di halaman sama
	Off page connector	Penghubung bagian flowchart di halaman berbeda

2.7.2 DFD

Diagram Alir Data (DAD) atau ***Data Flow Diagram*** **(DFD)** adalah suatu diagram yang menggunakan notasi-notasi untuk menggambarkan arus dari data sistem, yang penggunaannya sangat membantu untuk memahami sistem secara logika, terstruktur dan jelas. DFD merupakan alat bantu dalam menggambarkan atau menjelaskan DFD ini sering disebut juga dengan nama *Bubble chart*, *Bubble diagram*, model proses, diagram alur kerja, atau model fungsi.

DFD didesain untuk menunjukkan sebuah sistem yang terbagi-bagi menjadi suatu bagian sub-sistem yang lebih kecil dan untuk menggaris bawahi arus data antara kedua hal yang tersebut diatas. Diagram ini lalu “dikembangkan” untuk melihat lebih rinci sehingga dapat terlihat model-model yang terdapat di dalamnya.

Tabel 2.3 tabel simbol DFD

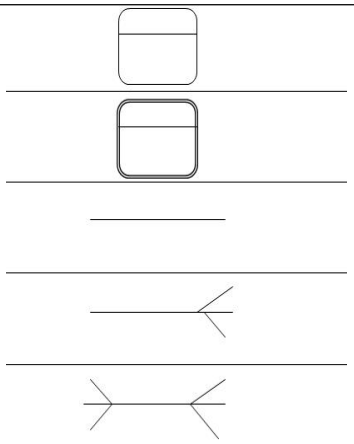
Simbol	Nama	Keterangan
	External Entity	Kesatuan di lingkungan luar sistem yang bisa berupa orang, organisasi, atau sistem lain
	Data Flow	Arus data ini mengalir diantara proses, simpan data dan kesatuan luar
	Data Store	Suatu file atau database pada sistem komputer atau catatan manual
	Proses	Proses seperti perhitungan aritmatik penulisan atau pembuatan laporan

Dari penjelasan tabel diatas DFD (Data Flow Diagram) mempunyai komponen / simbol yang memiliki bentuk dan kegunaan masing masing. Dan berguna untuk mengambarkan sebuah sistem yang akan dibentuk atau sistem yang akan di jalankan.

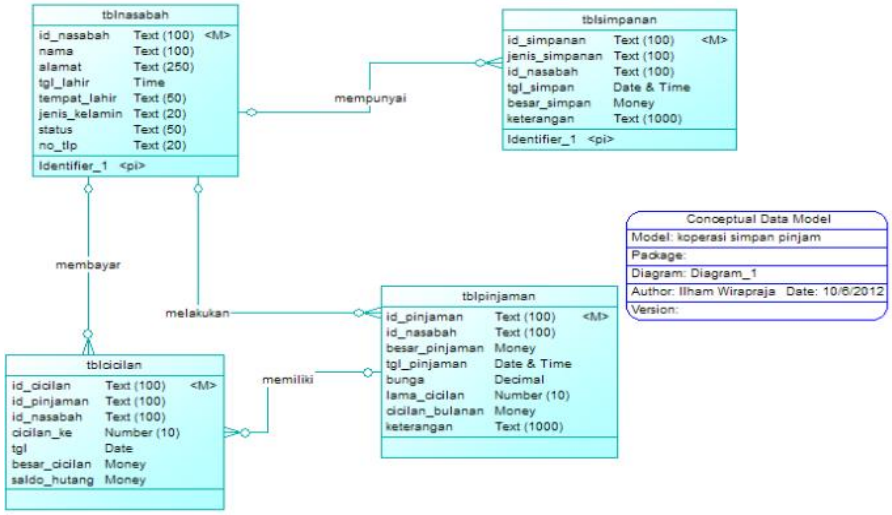
2.7.3 CDM

CDM singkatan dari Conseptual Data Model. CDM dipakai untuk menggambarkan secara detail struktur basis data dalam bentuk logik. Struktur ini independen terhadap semua software maupun struktur data storage tertentu yang digunakan dalam aplikasi ini. CDM terdiri dari objek yang tidak diimplementasikan secara langsung kedalam basis data yang sesungguhnya.

Tabel 2.4 Tabel Simbol CDM

Simbol	Keterangan
	Tabel View Relasi one to one Relasi one to many Relasi many to one

Dari penjelasan tabel diatas CDM (Conseptual Data Model) mempunyai komponen / simbol yang memiliki bentuk dan kegunaan masing masing. Dan berguna untuk mengambarkan sebuah sistem yang akan dibentuk atau sistem yang akan di jalankan. Untuk lebih jelasnya perhatikan gambar dibawah ini :

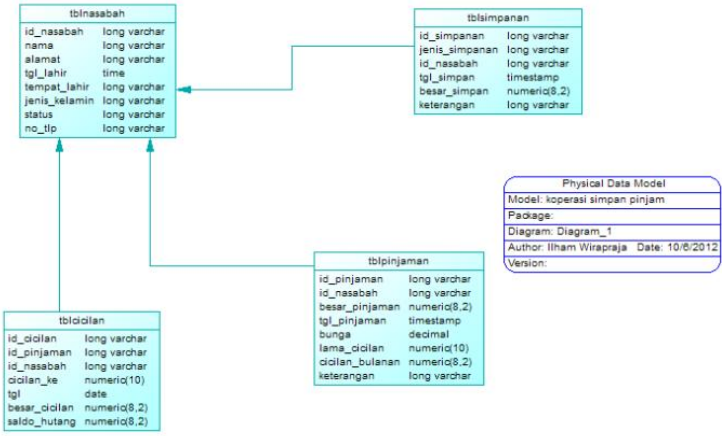


Gambar 2.2 Contoh CDM

2.7.4 PDM

PDM kependekan dari **Physical Data Model**. PDM merupakan gambaran secara detail basis data dalam bentuk fisik. Penggambaran rancangan PDM memperlihatkan struktur penyimpanan data yang benar pada basis data yang digunakan sesungguhnya. Physical Data Model atau yang biasa disebut PDM. PDM merupakan representasi fisik dari database yang akan dibuat dengan

mempertimbangkan DBMS yang akan digunakan. PDM dapat dihasilkan (di-generate) dari CDM yang valid.



Gambar 2.2 Contoh PDM

2.8 Apache HTTP Server

Apache HTTP Server merupakan aplikasi untuk server web terpopuler di dunia. Apache yang dipaketkan oleh XAMPP ini, sudah terdapat dua modul pengolah pemrograman di sisi server (server-side scripting), yaitu PHP dan Perl. Hal ini memungkinkan kita memanfaatkan web server untuk menginstall beberapa aplikasi berbasis web, atau untuk mempelajari pembuatan website dinamis menggunakan bahasa pemrograman tersebut di server lokal.

2.9 MySQL Database Server

Sebagaimana disebutkan sebelumnya, Apache memberikan kemampuan sebuah web server pada komputer kita, dan PHP memungkinkan kita menjalankan sebuah website dinamis yang menggunakan bahasa pemrograman PHP. Namun aplikasi berbasis web tidak bisa diinstall jika kita belum menyiapkan sebuah database server atau server basis data yang sesuai. Database server dibutuhkan untuk menyediakan penyimpanan data secara terstruktur, efektif, dan efisien. Bahkan banyak website besar dengan trafik yang tinggi memanfaatkan MySQL untuk penyimpanan basis datanya. Sebut saja Flickr, Facebook, Wikipedia, Google, Nokia dan YouTube yang secara resmi telah membeberkan bahwa website mereka menggunakan MySQL sebagai database server.

2.10 MySQL

MySQL adalah sebuah perangkat lunak sistem manajemen basis data SQL (bahasa Inggris: database management system) atau DBMS yang multithread, multi-user, dengan sekitar 6 juta instalasi di seluruh dunia. MySQL AB membuat MySQL tersedia sebagai perangkat lunak gratis dibawah lisensi GNU General Public License (GPL), tetapi mereka juga menjual dibawah lisensi komersial untuk kasus-kasus dimana penggunaannya tidak cocok dengan penggunaan GPL MySQL memiliki beberapa keistimewaan, antara lain :

1. **Portabilitas**. MySQL dapat berjalan stabil pada berbagai sistem operasi seperti Windows, Linux, FreeBSD, Mac Os X Server, Solaris, Amiga, dan masih banyak lagi.
2. Open Source. MySQL didistribusikan secara *open source*, dibawah lisensi GPL sehingga dapat digunakan secara cuma-cuma.
3. **Multiuser.** MySQL dapat digunakan oleh beberapa user dalam waktu yang bersamaan tanpa mengalami masalah atau konflik.
4. **Performance tuning**. MySQL memiliki kecepatan yang menakjubkan dalam menangani query sederhana, dengan kata lain dapat memproses lebih banyak SQL per satuan waktu.
5. **Jenis Kolom**. MySQL memiliki tipe kolom yang sangat kompleks, seperti signed / unsigned integer, float, double, char, text, date, timestamp, dan lain-lain.
6. **Perintah dan Fungsi.** MySQL memiliki operator dan fungsi secara penuh yang mendukung perintah Select dan Where dalam perintah (*query*).
7. **Keamanan.** MySQL memiliki beberapa lapisan sekuritas seperti level subnetmask, nama host, dan izin akses *user* dengan sistem perizinan yang mendetail serta sandi terenkripsi.
8. **Skalabilitas dan Pembatasan.** MySQL mampu menangani basis data dalam skala besar, dengan jumlah rekaman (*records*) lebih dari 50 juta dan 60 ribu tabel serta 5 milyar baris. Selain itu batas indeks yang dapat ditampung mencapai 32 indeks pada tiap tabelnya.

BAB III
METODOLOGI

Metode penelitian digunakan untuk menjelaskan dan memberikan gambaran tentang penelitian yang akan digunakan, terdiri dari *dataset* yang akan digunakan, daftar *class*, alur sistem, metode yang digunakan dan pembahasan tahapan *pre-processing* yang dilakukan.

3.1 Dataset

Pada penelitian ini akan digunakan 2 macam dataset, yaitu dataset pertama merupakan dataset yang terdiri dari 600 data *training*, dan data *testing* yang akan diujikan merupakan bagian dari data *training*, pada data *training* pertama akan dilakukan sejumlah pengujian dengan menggunakan jumlah data *testing* yang berbeda, yaitu 10, 25, 50 dan 100 data *testing*, yang dipilih secara acak oleh sistem pada setiap ujicobanya. Dataset pertama digunakan dengan tujuan untuk mengukur sejauh mana metode dapat mengenali kembali data yang sudah ada pada data *training*.

Selanjutnya, dataset kedua merupakan dataset yang terdiri dari dari 600 data training, seperti halnya dengan dataset pertama, perbedaan nya terletak pada data testing yang digunakan. Data testing pada dataset kedua bukan merupakan bagian dari data training, data testing yang akan diujikan berjumlah 100 soal baru. Dataset kedua digunakan untuk mengetahui kemapuan metode untuk mengkategorikan data soal baru.

Data yang digunakan merupakan data soal ujian biologi tingkat SMA yang diperoleh dari buku latihan soal, Ulangan Harian, modul biologi dan soal-soal lainnya dari kelas X sampai kelas XII.

3.2 Daftar Class

Daftar class merupakan hal penting yang dibutuhkan dalam sebuah penelitian yang bertujuan untuk mengategorikan atau mengklasifikasi sebuah data. Enam level yang ada pada *Cognitive domain* taksonomi bloom yang telah direvisi, akan dijadikan acuan pengkategorian data soal ujian pada penelitian ini. Daftar

class yang disesuaikan dengan *cognitive domain* revised taksonomi bloom dapat dilihat pada gambar 3.1

Gambar 3.1. Piramida level cognitive domain

Gambar 3.1 merupakan piramida yang menggambarkan susunan keenam level dari cognitive domain taksonomi bloom, tiga teratas disebut *Higher Order Thinking Skill*, sedangkan tiga level terbawah disebut *Lower Order Thinking Skill*. Jika dilihat, piramida gambar 3.1 semakin keatas semakin mengerucut, hal itu mencerminkan bahwa semakin sulit level soal tersebut, kemunculan nya di dalam daftar soal semakin sedikit. Setiap level memiliki ciri khas tersendiri dan memiliki daftar kata kerja operasional pada tabel 3.1 untuk membantu mencerminkan maksud dari setiap soal. Kata kerja operasional bukan merupakan aturan pasti dalam mengkategorikan soal ujian, melainkan untuk memudahkan menyampaikan maksud dari setiap level *cognitive domain* tersebut.

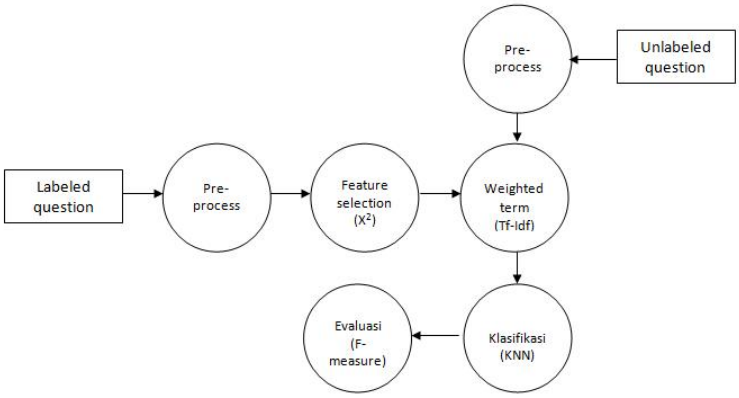
Tabel 3.1 : daftar class dan contoh kata kerja operasional

Kategori	Contoh Kata Kerja opersasiomal
mengingat	Mengutip, menyebutkan, mendaftar, menunujukkan, mengidentifikasi, melabeli, memasangkan, menamai, menandai, membaca, menyadari, menghafal, mencatat, mengulang memilih, menulis.
memahami	Menerangkan, menjelaskan, menterjemahkan,menguraikan, mengartikan, menyatakan kembali, menafsirkan, menginterpretasikan, mendiskusikan,menyeleksi, mendeteksi, melaporkan, menduga,mengelompokkan, memberi

menerapkan	Memilih,menerapkan, melaksanakan,mengubah, menggunakan, mendemonstrasikan, memodifikasi, menginterpretasikan, menunjukkan,membuktikan, menggambarkan, mengoperasikan, menjalankan, memprogramkan, mempraktekkan, memulai
menganalisis	Mengkaji ulang, membedakan, membandingkan, mengkontraskan, memisahkan, menghubungkan, menunjukan hubungan antara variabel, memecah menjadi beberapa bagian, menyisihkan, menduga,mempertimbangkan
mengevaluasi	Mengkaji ulang, mempertahankan, menyeleksi,mempertahankan, mengevaluasi,mendukung, menilai, menjustifikasi,mengecek, mengkritik, memprediksi, membenarkan, menyalahkan.
membuat	Merakit, merancang, menemukan, menciptakan, memperoleh, mengembangkan,memformulasikan, membangun,membentuk, melengkapi,membuat, menyempurnakan, melakukan, inovasi, mendisain, menghasilkan karya.

3.3 Alur Sistem

Pada penelitian ini akan digunakan 2 jenis data, yaitu *unlabeled data* dan *labeled data*. *Labeled data* akan digunakan sebagai data *training*. Alur sistem pada penelitian ini akan ditunjukkan pda gambar 3.2.



Gambar 3.2. Alur sistem

Seperti yang ditunjukkan oleh alur sistem pada gambar 3.2, algoritma yang digunakan adalah algoritma K-Nearest Neighbour, sebelum masuk ke tahap klasifikasi dilakukan pembobotan menggunakan Tf-Idf dan seleski fitur

menggunakan metode Chi-Square. Metode Chi-square hanya bisa diterapkan pada labeled data saja.

3.4 Tf-Idf

TF-IDF (Term Frequency-Inverse Document Frequency) merupakan metode yang biasa dipakai dalam ranah *text mining*, dan digunakan untuk memberi bobot pada *term* sebagai strategi untuk mengklasifikasikan *text* atau dokumen

Proses pemberian bobot pada term dengan menggunakan TF-IDF, terdiri dari beberapa proses, yaitu menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap *term* dalam setiap data yang akan diberi bobot. Berikut adalah formula yang digunakan untuk bobot term (t) yang ada pada dokumen (d), yaitu :

$$w_{t,d} = tf_{t,d} \times idf_t \quad (1)$$

tf merupakan variabel yang menunjukkan jumlah term (t) dalam dokumen (d). Sedangkan *idf* merupakan variabel yang menunjukkan jumlah term (t) pada seluruh dokumen yang ada pada data, berikut adalah formula yang digunakan untuk mendapatkan nilai *idf* :

$$idf_t = \log \frac{N}{df_t} \quad (2)$$

df_t pada formula (2) merupakan jumlah dokumen yang di dalamnya terdapat term *t*, kemudian variabel *N* merupakan jumlah seluruh dokumen yang digunakan dalam data.

3.5 Chi-Square

Chi-Square merupakan salah satu metode feature selection yang termasuk metode filter. Metode feature selection terbukti dapat meningkatkan akurasi pada beberapa penelitian sebelumnya. Berikut adalah formula yang digunakan untuk menerapkan metode *feature selection* chi-square.

$$X^2(t,c) = \frac{N(A \times D - B \times C)^2}{(A+B) \times (C+D) \times (A+C) \times (B+D)} \quad (3)$$

variabel t merupakan term yang dicari pada sebuah class c, N merupakan jumlah dokumen *training*, A merupakan jumlah dokumen pada kelas c yang mengandung *term* t, B merupakan jumlah dokumen yang tidak ditemukan pada *class* c tapi mengandung *term* t, C adalah jumlah dokumen yang ditemukan di *class* c namun tidak mengandung *term* t , dan D adalah jumlah dokumen yang bukan merupakan dokumen kelas c dan tidak mengandung *term* t. Variabel-variabel tersebut didapatkan dari korelasi term dan class yang didapatkan dari tabel kontingensi pada tabel 3.2

Tabel 3.2. Tabel Kontingensi Term dan Class

	t	not t
C	A	C
not C	B	D

3.6 K-Nearest Neighbour

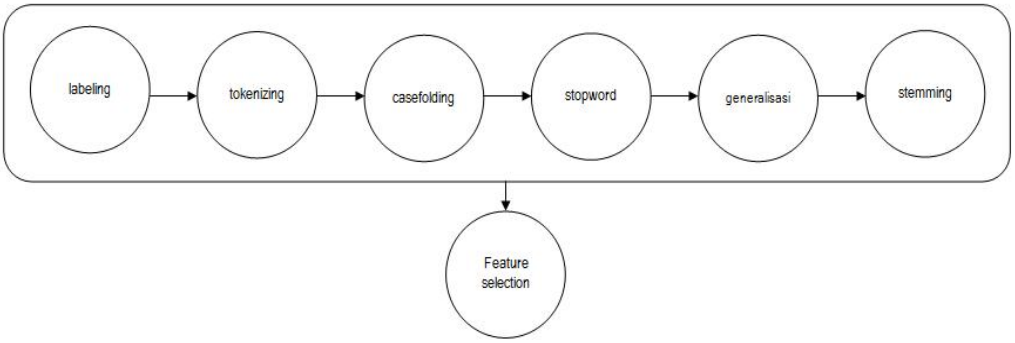
K-NN (K-nearest neighbour) merupakan classifier yang digunakan untuk mengklasifikasikan data berdasarkan perbandingan nilai K tetangga terdekat, parameter K pada KNN memiliki pengaruh besar pada hasil akhir prediksi yang dihasilkan. KNN mengukur similarity data menggunakan pengukuran distance. Pada penelitian ini digunakan *Euclidean distance* untuk menghitung *similarity* dari dua data atau lebih.

$$d_{(x_1,x_2)} = \sqrt{(|x_{11} - x_{21}|)^2 + (|x_{12} - x_{22}|)^2} \qquad (4)$$

variabel d merupakan *distance* dari dua data. KNN dikenal sebagai metode yang membutuhkan waktu dan cost komputasi yang besar, karena harus menyocokkan data tes, dengan seluruh data *training* yang ada. Sehingga dibutuhkan adanya proses *feature selection* untuk membuang fitur yang tidak dibutuhkan dalam data, sehingga diharapkan mampu mempercepat waktu komputasi.

3.7 Pre-Processing

Sebelum dilakukan tahapan kategori yang sesungguhnya, data yang digunakan harus melalui tahapan *pre-processing*. Pada penelitian ini tahapan *preprocessing* terdiri dari labelling yang dilakukan secara semi-otomatis, sedangkan *tokenizing, casefolding, stopword* dan *stemming* dilakukan otomatis oleh sistem.



Gambar.3.3 tahapan pre-processing

Pada gambar 3.3 dijelaskan tahapan preprocess secara garis besar yang dilakukan pada penelitian ini, "FS" merupakan tahapan *preprocessing* ketika metode *Feature Selection* Chi-Square diterapkan setelah proses *stemming* selesai dilakukan.

3.8 Evaluasi

Tahap evaluasi merupakan sebuah tahapan yang dilakukan setelah proses ujicoba selesai dilakukan, tahapan evaluasi digunakan untuk mengukur dan mengetahui kinerja dari metode yang sedang diuji. Pada penelitian ini akan digunakan metode *F-measure* untuk mengetahui kinerja dari KNN dan metode *feature selection* Chi-Square.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (6)$$

$$\text{F} - \text{measure} = \frac{2\text{PR}}{\text{P+R}} \quad (7)$$

parameter TN,TP ,FP ,FN didapatkan dari tabel *confusion matrix* pada tabel 3.3.

Tabel 3.3. Tabel confusion matrix

		Prediksi	
		Ya	Tidak
aktual	Ya	TP	FP
	Tidak	FN	TN

BAB IV
HASIL DAN PEMBAHASAN

Implementasi sistem menggunakan bahasa pemograman PHP, berbasis web. Fitur yang ada pada sistem meliputi preprocessing, hitung Chi-Square, seleksi fitur, term weighting, hitung similarity menggunakan euclidean distance, dan klasifikasi dokumen menggunakan K-Nearest Neighbour. Output sistem merupakan hasil klasifikasi soal menggunakan KNN dan metode feature selection Chi-Square, dapat dilihat pada gambar 4.1

Gambar 4.1 Output Sistem

Sebelum memulai proses pemilihan fitur, yang dilanjutkan dengan tahapan klasifikasi, data soal akan di pre-processing terlebih dahulu, contoh data soal yang akan di pre-processing bisa dilihat pada tabel 4.1.

Tabel 4.1. Data soal

id_soal	soal	class
1	apa saja ciri-ciri jamur Oomycotina	C1
2	Apa yang dimaksud dengan gangguan pernapasan asfiksia	C1
3	apakah yang dimaksud dengan kolkisin?	C1
4	Faktor-faktor apakah yang dapat menurunkan kekebalan tubuh	C1
5	Jelaskan perbedaan kemosintesis dengan fotosintesis.	C2

Dengan menggunakan tabel kontingensi pada tabel 3.2, hitung variabel A, B, C dan D pada setiap term soal, sehingga didapatkan hasil pada tabel 4.2

Tabel 4.2.Perhitungan Korelasi

term	C1				C2				C3				C4				C5				C6			
	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
apa	31	72	47	218	15	88	63	202	2	101	27	238	15	88	12	253	29	74	38	227	11	92	77	188
saja	7	2	93	498	2	7	98	493	0	9	100	491	0	9	98	493	0	9	100	491	0	9	100	491
ciri	8	0	95	500	0	8	100	495	0	8	100	495	0	8	98	497	0	8	100	495	0	8	100	495
jamur	1	0	99	500	0	1	100	499	0	1	100	499	0	1	98	501	0	1	100	499	0	1	100	499
xxx	70	764	43	145	126	708	36	152	209	625	18	170	122	712	27	161	152	682	32	156	152	682	32	156
apakah	17	35	83	466	17	35	83	466	0	52	100	449	2	50	96	453	16	36	85	464	0	52	100	449
maksud	25	1	75	499	1	25	99	475	0	26	100	474	0	26	98	476	0	26	100	474	0	26	100	474
belah	1	11	99	490	6	6	95	494	0	12	100	489	0	12	98	491	2	10	98	491	3	9	97	492
jelaskan	3	64	97	423	55	12	45	475	0	67	100	420	1	66	97	423	5	62	86	434	3	64	93	427
beda	1	46	99	453	41	6	59	493	0	47	100	452	1	46	97	455	1	46	98	454	3	44	97	455

Selanjutnya dengan menggunakan formula 3, didapatkan hasil perhitungan chi-square pada tabel 4.3 sebagai berikut

Tabel 4.3. Hasil Chi-square

term	C1	C2	C3	C4	C5	C6
apa	11.06	6.14	11.32	17.91	15.50	22.44
saja	24.56	0.20	1.82	1.78	1.82	1.82
ciri	39.16	1.60	1.60	1.56	1.60	1.60
jamur	5.008	0.20	0.20	0.19	0.20	0.20
xxx	19.20	1.10	12.50	0.005	0.08	0.08
maksud	123.63	3.21	5.43	5.30	5.43	5.43
ganggu	0.61	0.61	0.61	0.63	0.61	0.61
napas	0.20	9.73	0.80	0.78	0.80	0.80
apakah	10.55	10.55	11.34	6.46	7.92	11.34
faktor	41.71	3.39	4.23	0.25	1.42	1.42
dapat	5.78	6.38	11.21	9.84	26.80	0.19
turun	1.64	3.71	10.31	1.77	0	1.65
kebal	0.20	0.20	0.80	0.22	0.81	0.20
jelaskan	8.62	231.48	15.87	12.84	3.81	7.97
beda	7.79	182.77	10.23	7.56	7.67	3.90

Berdasarkan langkah perhitungan yang dilakukan secara manual diatas juga diterapkan pada implementasi sistem, sehingga contoh data soal pada tabel 4.1 setelah menggunakan metode feature selection akan menjadi seperti pada tabel 4.4. Proses pemilihan feature tergantung pada penetapan nilai *threshold* yang telah ditetapkan, yaitu 0,5

Tabel 4.4. Hasil akhir Chi-square

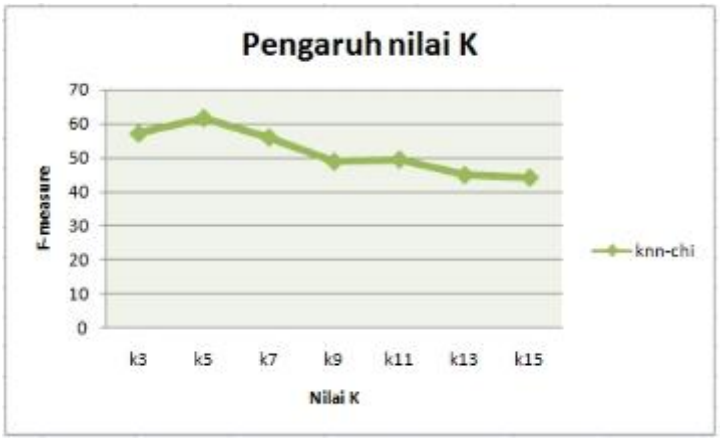
Soal	Isi Soal	Kategori
1	apa saja ciri ciri jamur xxx	C1
2	apa maksud ganggu xxx	C1
3	apakah maksud xxx	C1
4	faktor faktor apakah dapat turun	C1
5	jelaskan beda xxx xxx	C2

Penelitian ini dilakukan dengan tujuan untuk mengategorikan data soal ujian ke dalam 6 level *cognitive domain* taksonomi bloom, dengan menggunakan algoritma K-Nearest Neighbour (KNN) dan metode *feature selection* Chi-Square (X^2), sekaligus mengetahui pengaruh penggunaan metode *feature selection* terhadap perubahan perfomance algoritma KNN. Pengkategorian dengan menggunakan KNN tergantung pada nilai K yang telah ditentukan. Nilai K yang akan diujicobakan pada dataset kedua, dijelaskan pada tabel 4.5.

Tabel 4.5. Nilai K pada KNN-chi

k	KNN-chi
k=3	56,99
k=5	61,56
k=7	55,9
k=9	48,71
K=11	49,36
K=13	44,76
k=15	44,01

Berdasarkan pada tabel 4.5, pemilihan nilai K berpengaruh pada perfomance KNN, semakin besar nilai K yang ditentukan, batas-batas antar *class* semakin kabur sehingga menyebabkan *perfomance* menurun. Sehingga, dipilih nilai K=5 untuk ujicoba selanjutnya, untuk memudahkan pengamatan perubahan pengaruh nilai K terhadap nilai *F-measure* KNN, tabel 3 ditampilkan dalam grafik pada gambar 4.2.



Gambar 4.2 pengaruh nilai K terhadap F-measure

Selanjutnya, ujicoba masih menggunakan dataset kedua dan bertujuan untuk mengetahui pengaruh kombinasi data *training* terhadap 100 data *testing* baru yang bukan merupakan bagian dari data training. Rincian kombinasi data *training* dan hasil nilai *F-measure* untuk setiap ujicoba yang dilakukan, dijelaskan pada tabel 4.6

Tabel 4.6. Kombinasi data training

%	Jumlah data training	KNN	KNN-chi	Selisih
25%	150	18,52	41,65	23,13
50%	300	44,28	46,79	2,51
75%	450	42,67	48,07	5,4
100%	600	44,75	61,56	16,81

Berdasarkan pada tabel 4.2, metode *feature selection* Chi-square mampu meningkatkan nilai F-measure pada KNN, dengan menggunakan 600 data training peningkatannya cukup signifikan dengan perbandingan selisih antara

KNN tanpa feature selection dan KNN dengan menggunakan metode *feature selection* Chi-square mencapai16,81. Berdasarkan pada ujicoba tabel 4 jumlah data training sangat berpengaruh pada *perfomance* metode, semakin besar jumlah data training semakin tinggi nilai *F-measure* yang diperoleh.

Ujicoba selanjutnya, menggunakan dataset pertama dengan menggunakan data *testing* yang merupakan bagian dari data *training* dengan jumlah 10, 25, 50, dan 100. Hasil ujicoba menggunakan dataset pertama dijelaskan pada tabel 4.7.

Tabel 4.7. Hasil ujicoba dataset pertama

Jumlah data testing	KNN-chi
10	42,34
25	79,36
50	74,41
100	76,71

Pada beberapa kali percobaan yang dilakukan, metode *feature selection* chi-square berhasil meningkatkan *perfomance* KNN, fungsi chi-square yaitu menghapus fitur yang tidak diperlukan dan hanya menggunakan fitur yang memenuhi batas *threshold* yang telah ditentukan. Pada penelitian ini *threshold* yang digunakan untuk *Chi-square* adalah 0,5.

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan ujicoba yang telah dilakukan, KNN dan metode *feature selection* chi-square berhasil diterapkan untuk mengkategorikan soal ujian biologi tingkat SMA berdasarkan *cognitive domain* taksonomi bloom, dengan nilai *F-measure* tertinggi pada dataset pertama mencapai 79,36%, dan nilai F-measure tertinggi pada dataset kedua mencapai 61,56%

Metode feature selection Chi-Square terbukti mampu meningkatkan *perfomance* algoritma K-nearest neighbour hingga 16,81 pada dataset kedua. Jumlah kombinasi data *training* memiliki pengaruh besar pada *perfomance* metode.

5.2 Saran

Dari banyaknya kelemahan dan kekurangan yang terdapat pada aplikasi ini, maka bagi para pembaca yang ingin mengembangkan Pengkategorian Soal Ujian Berdasarkan Cognitive Domain Taksonomi Bloom, disarankan untuk :

1. Menambah Data dan menambah perbandingan metode klasifikasi, Karena ilmu selalu berkembang

DAFTAR PUSTAKA

D. A. Abduljabbar and N. Omar, “Exam Questions Classification Based On Bloom’s Taxonomy Cognitive Level Using Classifiers Combination,” *J. Theor. Appl. Inf. Technol.*, vol. 78, no. 3, pp. 447–455, 2015.

D. Zhu and J. Xiao, “R-tfidf , a Variety of tf-idf Term Weighting Strategy in Document Categorization,” *Seventh Int. Conf. Semant. Knowl. Grids R-tfidf,* pp. 83–90, 2011.

F. Thabtah, M. A. Eljinini, M. Zamzeer, and W. M. Hadi, “Naïve Bayesian Based on Chi Square to Categorize Arabic Data,” *Commun. IBIMA*, vol. 10, pp. 1943–7765, 2009.

J. Ling and T. B. Oka, “Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square,” vol. 3, no. 3, pp. 92–99, 2014.

R. Utari, W. Madya, and Pusdiklat KNPK, “Taksonomi bloom Apa dan Bagaimana Menggunakannya?”

Y. Yang and J. Pedersen, “A comparative Study on Feature Selection in Text Categorization,” 1997.