



## Feature Selection Chi-Square dan K-NN pada Pengkategorian Soal Ujian Berdasarkan *Cognitive Domain* Taksonomi Bloom

Indah Listiowarni<sup>1</sup>, Eka Rahayu Setyaningsih<sup>2</sup>

<sup>1</sup>Sekolah Tinggi Teknik Surabaya, email : listioindah@gmail.com

<sup>2</sup>Sekolah Tinggi Teknik Surabaya, email : eka@stts.edu

### 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 ( $X^2$ ) 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

### Abstract

*Examination is the final benchmark for learners with the aim to identify the cognitive and capability of the students during the teaching and learning activities in the classroom, it need the right set of question that required during the exam. The manual data classification process is very long and complicated, because it also deals with the amount of data and checks each term, so it takes a system that can automatically be applied to make the process of classifying data more easily and not spending much time. Taxonomy bloom is a taxonomy used in the world of education contains six levels of cognitive based on the level of difficulty. This study concentrates on categorizing data on high school biology tests based on cognitive domains of taxonomic bloom, and K-Nearest Neighbour (KNN), and feature selection method Chi-Square ( $X^2$ ) aimed at selecting required features . The dataset used in this study consists of 2 kinds of datasets, the first dataset method used successfully achieved the highest F-measure value 79.36%, while in the second dataset, managed to achieve the highest F-measure value 61.56%*

**Keywords :** KNN, bloom's taxonomy, Chi-Square, F-measure

## 1. Pendahuluan

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 penyusunan 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*) [1].

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 algoritma yang 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 menyeleksi fitur yang dianggap penting dan membuang fitur-fitur yang tidak diperlukan. Pada penelitian sebelumnya, K-Nearest Neighbour memiliki *performance* terbaik ketika digunakan bersama metode *feature selection*, dibandingkan dengan dua classifier pembandingnya, yaitu Naive Bayesian (NB) dan Support Vector Machine (SVM) [2].

## 2. Metode Penelitian

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.

### 2.1 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 600 data training, seperti halnya dengan dataset pertama, perbedaannya 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 kemampuan 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.

## 2.2 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 1 [1].



**Gambar 1. Piramida level cognitive domain**

Gambar 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 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 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.

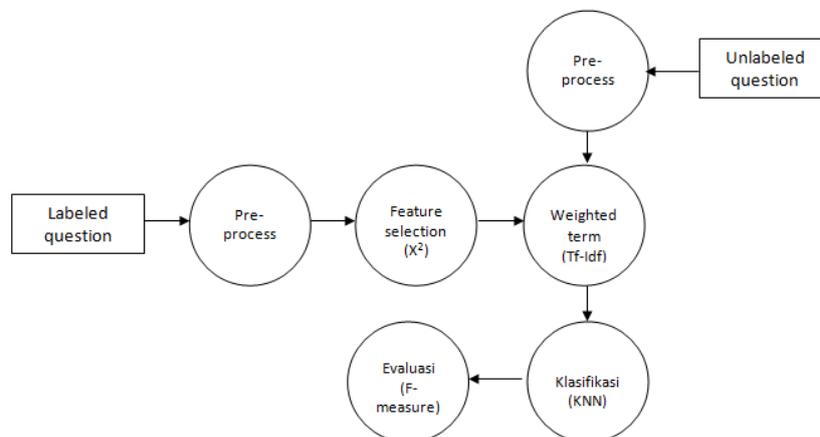
**Tabel 1 : daftar class dan contoh kata kerja operasional [1]**

Kategori	Contoh Kata Kerja opsasional
mengingat	Mengutip, menyebutkan, mendaftar, menunjukkan, mengidentifikasi, melabeli, memasang, 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, menunjukkan 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.

### 2.3 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 pada gambar 2.



**Gambar 2. Alur sistem**

Seperti yang ditunjukkan oleh alur sistem pada gambar 2, algoritma yang digunakan adalah algoritma K-Nearest Neighbour, sebelum masuk ke tahap klasifikasi dilakukan pembobotan menggunakan Tf-Idf dan seleksi fitur menggunakan metode Chi-Square. Metode Chi-square hanya bisa diterapkan pada labeled data saja.

## 2.4 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 [3]

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.

## 2.5 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 [2][4][5]. 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*  $t$  [2]. Variabel-variabel tersebut didapatkan dari korelasi term dan class yang didapatkan dari tabel kontingensi pada tabel 2

**Tabel 2. Tabel Kontingensi Term dan Class**

	t	not t
C	A	C
not C	B	D

## 2.6 K-Nearest Neighbour

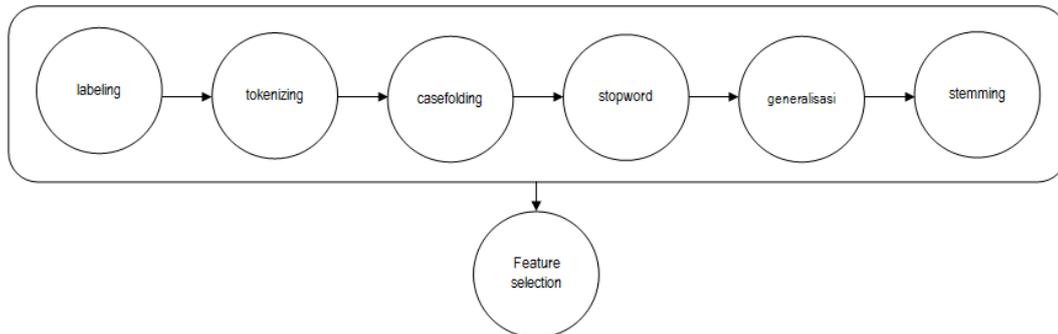
K-NN (K-nearest neighbour) merupakan classifier yang digunakan untuk mengklasifikasikan data berdasarkan perbandingan nilai K tetangga terdekat, parameter K pada KNN memiliki pengaruh besar pada hasil akhir prediksi yang dihasilkan. KNN mengukur similarity data menggunakan pengukuran distance. Pada penelitian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

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

variabel  $d$  merupakan *distance* dari dua data. KNN dikenal sebagai metode yang membutuhkan waktu dan cost komputasi yang besar, karena harus mencocokkan 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 [6].

## 2.7 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** tahapan pre-processing

Pada gambar 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.

## 2.8 Evaluasi

Tahap evaluasi merupakan sebuah tahapan yang dilakukan setelah proses uji coba 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.

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

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

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

parameter TN,TP ,FP ,FN didapatkan dari tabel *confusion matrix* pada tabel 2.

**Tabel 2.** Tabel confusion matrix

		Prediksi	
		Ya	Tidak
aktual	Ya	TP	TN
	Tidak	FP	FN

### 3. Implementasi

Implementasi sistem menggunakan bahasa pemrograman 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

no	soal	hasil
1	apa maksud teori and xxx menurut	C1
2	sebutkan faktor faktor apa dapat halang langsung proses	C1
3	apa maksud xxx	C1
4	sebutkan macam macam kandung xxx	C1
5	apakah fungsi variabel kontrol variabel bebas teliti	C1
6	sebutkan ciri ciri xxx xxx	C1
7	apa maksud belah xxx	C1
8	apakah fungsi manfaat xxx	C1
9	sebutkan macam macam modifikasi fungsi tumbuh	C1
10	sebutkan masuk tunjang fungsi	C1
11	sebutkan macam macam fungsi	C1
12	sebutkan faktor faktor cetus xxx	C1
13	sebutkan bagi bagi apa anatomi xxx cermin pola segmentasi xxx	C1
14	sebutkan asumsi estimasi ukur xxx metode	C1
15	apakah xxx hasil bilus	C1
16	apakah xxx bagi xxx	C1
17	sebutkan zona tumbuh kembang	C1
18	sebutkan faktor faktor pengaruh tumbuh	C1
19	sebutkan fungsi xxx tumbuh	C1
20	apakah pengaruh tumbuh	C1
21	sebutkan ciri xxx	C1
22	apa maksud xxx	C1
23	sebutkan macam hasil	C1

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 3.

**Tabel 3. 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 2, hitung variabel A, B, C dan D pada setiap term soal, sehingga didapatkan hasil pada tabel 4

**Tabel 4.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 5 sebagai berikut

**Tabel 5. 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 3 setelah menggunakan metode feature selection akan menjadi seperti pada tabel 6. Proses pemilihan feature tergantung pada penetapan nilai *threshold* yang telah ditetapkan, yaitu 0,5

**Tabel 6. 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

#### 4. Hasil dan Pembahasan

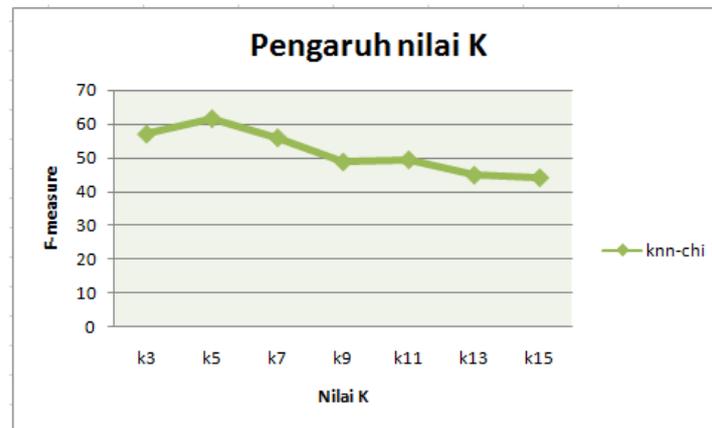
Penelitian ini dilakukan dengan tujuan untuk mengategorikan data soal ujian ke dalam 6 level *cognitive domain* taksonomi bloom, dengan menggunakan algoritma K-Nearest Neighbour (KNN) dan metode *feature selection* Chi-Square ( $X^2$ ), sekaligus mengetahui pengaruh penggunaan metode *feature selection* terhadap perubahan performance algoritma KNN. Pengkategorian dengan menggunakan KNN tergantung pada nilai K yang telah ditentukan. Nilai K yang akan diujicobakan pada dataset kedua, dijelaskan pada tabel 7.

**Tabel 7. 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 7, pemilihan nilai K berpengaruh pada performance KNN, semakin besar nilai K yang ditentukan, batas-batas antar *class* semakin kabur sehingga menyebabkan *performance* 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 5.



**Gambar.5 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 8

**Tabel 8. 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, 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 mencapai 16,81. Berdasarkan pada ujicoba tabel 4 jumlah data *training* sangat berpengaruh pada *performance* 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 9.

**Tabel 9. 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 *performace* 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.

## 5. 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 *performace* algoritma K-nearest neighbour hingga 16,81 pada dataset kedua. Jumlah kombinasi data *training* memiliki pengaruh besar pada *performace* metode.

## 6. Referensi

- [1] R. Utari, W. Madya, and Pusdiklat KNPK, "Taksonomi bloom Apa dan Bagaimana Menggunakannya?"
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] Y. Yang and J. Pedersen, "A comparative Study on Feature Selection in Text Categorization," 1997.