# Prediksi Kelulusan Mahasiswa Menggunakan *Data mining* Algoritma *K-means*

**Ray Mondow Sagala**
Perguruan Advent II Bandung, Bandung, Indonesia
e-mail: raymondows123@gmail.com

## Abstrak

Kelulusan sebuah mata kuliah sangatlah penting, apabila terdapat mahasiswa yang tidak lulus di sebuah mata kuliah terutama mata kuliah yang memiliki keterikatan dengan mata kuliah lain, harus mengambil ulang mata kuliah tersebut. Kelulusan di sebuah mata kuliah tidak dapat diketahui sebelum dilakukannya ujian final dan perhitungan nilai akhir sebuah mata kuliah,. Untuk itu perlu dilakukannya prediksi terhadap kelulusan mata kuliah agar membantu mengantisipasi tidak lulusnya mahasiswa dalam sebuah mata kuliah. Melalui tahap studi literatur, wawancara dan melihat data kinerja akademik yang di dapat, maka digunakan nilai tugas, nilai *unit test*, nilai *mid test*, dan kehadiran yang didapat dari data kinerja serta faktor internal dan eksternal mahasiswa seperti keaktifan kelas, Status tinggal, Bahasa materi pelajaran, dan bentuk tugas akhir yang di berikan. Pengolahan data dilakukan menggunakan *K-means*, kemudian menghitung *chi square atribute* selection dan menghitung tingkat akurasi prediksi tersebut menggunakan confusion matrix. Hasil dari penelitian menunjukkan bahwa hasil prediksi menggunakan K = 3 dari 118 data yang diolah terdapat 13 mahasiswa yang tidak lulus, 36 mahasiswa lulus dengan nilai cukup, dan 69 mahasiswa lulus dengan nilai baik. Dan yang mempengaruhi prediksi kelulusan menggunakan *chi square atribute* adalah nilai *mid* dengan *ranked atributes* sebesar 49, nilai tugas sebesar 46, dan kehadiran sebesar 42. Bahasa materi yang menggunakan Bahasa inggris mempengaruhi mahasiswa yang lulus dengan nilai cukup. Sedangkan tugas akhir yang berupa proyek atau ujian teori tidak terlalu mempengaruhi prediksi kelulusan mahasiswa di sebuah mata kuliah. Penggunaan *confusion matrix* terhadap hasil prediksi menunjukkan tingkat akurasi sebesar 93% dengan presisi dan *recall* sebesar 96% dan 92%.

**Kata kunci**: Kelulusan, Mata Kuliah, Prediksi, *K-means*, *Chi square atribute Selection*, confusion matrix.

# *Prediction of college subject using K-means Algorithm in Data mining*

## *Abstract*

*Graduation of a course is very important, if there are students who do not pass a course, especially subjects that have an attachment to another course, must take back the course. Graduation in a course cannot be known before the final exam and final grade calculation are calculated. For this reason, it is necessary to predict the graduation of courses to help anticipate what makes students fail in a course.Through the literature study stage, interviews and looking at academic performance data obtained, the assignment values, unit test values, mid test scores, and attendance obtained from performance data and internal and external factors of students such as class activity, status of residence, language lesson, and the form of the final project given. Data processing is performed using K-means, then calculate the chi square atribute selection and calculate the accuracy of the prediction using a confusion matrix.The results of the study showed that the prediction results using K = 3 of 118 data processed there were 13 students who did not pass, 36 students graduated with sufficient grades, and 69 students graduated with good*

*grades. And what influences the graduation prediction using the chi square atribute is a mid value with ranked atributes of 49, an assignment value of 46, and attendance of 42. Language material using English influences students who graduate with sufficient grades. While the final project in the form of a project or a theory test does not greatly affect the predictions of student graduation in a course. The use of confusion matrix to the prediction results shows an accuracy rate of 93% with precision and recall of 96% and 92%.*

***Keywords**: Graduation, Course, Prediction, K-means, Chi square atribute Selection, confusion matrix**.***

## 1. Pendahuluan

Kelulusan di dalam sebuah mata kuliah adalah sebuah harapan besar bagi seorang mahasiswa karena terdapat beberapa mata kuliah yang memiliki keterikatan dengan mata kuliah lainnya, Sehingga kelulusan mahasiswa di sebuah mata kuliah dapat mempengaruhi pengambilan mata kuliah disemester yang akan datang. Misalnya, untuk mengambil mata kuliah algoritma pemrograman 2 di semester genap harus telah lulus dari mata kuliah algoritma pemrograman 1 di semester ganjil.

Terdapat beberapa mata kuliah yang memiliki keterikatan dengan mata kuliah lainnya, Sehingga kelulusan mahasiswa di sebuah mata kuliah dapat mempengaruhi pengambilan mata kuliah disemester yang akan datang. Mahasiswa yang lulus di sebuah mata kuliah tidak dapat diketahui sebelum dilakukannya perhitungan nilai akhir sebuah mata kuliah, sehingga, apabila terdapat mahasiswa yang tidak lulus di sebuah mata kuliah terutama mata kuliah di semester ganjil yang memiliki keterikatan dengan mata kuliah di semester genap, akan mengalami kesulitan untuk mengambil mata kuliah tersebut, karena mahasiswa itu harus melakukan pengambilan ulang mata kuliah yang gagal tersebut hanya di semester ganjil yang akan datang, jika hal ini sering terjadi, dapat membuat mahasiswa tersebut terlambat untuk diwisuda.

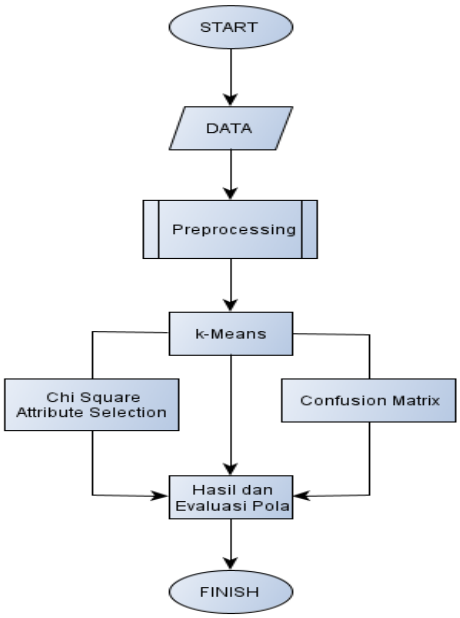
Maksud dilakukannya penelitian ini adalah supaya mahasiswa dapat mengetahui hal-hal apa saja yang dapat membuat seorang mahasiswa tidak lulus dari mata kuliah. Dan apa saja yang dapat diperhatikan dosen agar dapat mengetahui kinerja dan kemampuan mahasiswanya di kemudian hari. Tujuan penelitian ini adalah untuk mengetahui bagai mana cara mengolah data mahasiswa yang dapat digunakan untuk memprediksi kelulusan mahasiswa dalam sebuah mata kuliah menggunakan *data mining* algoritma *K-means*.

Berdasarkan studi literatur terhadap jurnal atau penelitian yang pernah dilakukan sebelumnya, dan dilakukannya wawancara, maka atribut yang digunakan pada penelitian ini adalah kehadiran, status tinggal, Bahasa, tugas akhir, nilai tugas, nilai tugas, nilai *unit test*, dan *mid test*, keaktifan kelas. Berdasarkan atribut di atas terdapat atribut non numerik karna Metode *K-means* adalah metode yang bisa di lakukan apabila data yang digunakan adalah data yang berupa angka[6]. maka proses *binning* perlu dilakukan. Proses *Binning* disebut juga proses normalisasi, yaitu proses untuk mentransformasi nilai-nilai dari data-data non-numerik menjadi data-data yang bisa dikalkulasi[6]. Aturan untuk proses *binning* dalam penelitian ini adalah sebagai berikut[8] :

1. Untuk atribut dengan dua kategori, setiap kategori di-set dengan biner 1 atau 0 sesuai dengan nilai atribut.
2. Untuk atribut dengan lebih dari dua kategori, nilai dari atribut itu akan bernilai biner dengan jumlah digit biner sesuai dengan jumlah kategori dalam atribut tersebut. Dan pengesetan digit 1 akan sesuai dengan kondisi true dari setiap kategori atribut.

## 2. Metode Penelitian

Sebelum pengolahan data dilakukan, terlebih dahulu dilakukan *preprocessing* pada data dan kemudian melakukan pengolahan data menggunakan langkah-langkah pada algoritma *K-means* untuk memprediksi mahasiswa yang lulus dari mata kuliah. kemudian melakukan seleksi atribut dari hasil prediksi menggunakan *Chi-Square Atribute Selection* untuk mengetahui atribut apa saja yang mempengaruhi kelulusan mahasiswa dalam sebuah mata kuliah, lalu mencari tingkat akurasi, presisi, dan *recall* dari hasil prediksi menggunakan *Confusion Matrix*.




**Gambar 1** *Flowchart* Pengolahan Data

### *Data mining*

Kata mining merupakan kiasan dari bahasa inggris, *mine*. Jika *mine* berarti menambang sumber daya yang tersembunyi di dalam tanah, maka *Data mining* merupakan penggalian makna yang tersembunyi dari kumpulan data[3]. Berikut ini tahapan proses *… data mining*:

1. Pembersihan Data, yaitu: menghapus data pengganggu (*noise*) dan mengisi data yang hilang.
2. Integrasi Data, yaitu: menggabungkan berbagai sumber data.
3. Pemilihan Data, yaitu: memilih data yang relevan.
4. Transformasi Data, yaitu: mentransformasi data ke dalam format untuk diproses dalam *data mining*.
5. Proses mining, yaitu: Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data
6. Evaluasi pola: :Untuk mengidentifikasi pola-pola ke dalam *knowledge based* yang ditemukan.

### *Algoritma K-means*

Algoritma *K-means* pada dasarnya melakukan dua proses, yakni proses pendeteksian lokasi pusat tiap *cluster* dan proses pencarian anggota dari tiap-tiap *cluster*[4]. *K-means Clustering* merupakan salah satu metode data *Clustering* non-hirarki yang mengelompokkan data dalam bentuk satu atau lebih *cluster*/kelompok. Data-data yang memiliki karakteristik yang sama dikelompokkan dalam satu *cluster*/kelompok dan data yang memiliki karakteristik yang berbeda dikelompokkan dengan *cluster*/kelompok yang lain sehingga data yang berada dalam satu *cluster*/kelompok memiliki tingkat variasi yang kecil[1]. Algoritma dasar dari *K-means Clustering* dapat ditentukan dengan langkah-langkah sebagai berikut$^{n}$ :

1. Menentukan jumlah *cluster* yang diinginkan.
2. Memilih *cluster* secara *random* dan mengelompokkan data yang lainnya ke dalam kluster-kluster tersebut berdasarkan jarak terdekatnya.
3. Menghitung *centroid*/ rata-rata dari data yang ada di hasilkan dari masing-masing *cluster*.
4. Mengalokasikan kembali masing-masing data ke dalam *centroid* / rata-rata kluster yang terdekat.
5. Ulangi langkah ke-3, apabila masih ditemukan data yang berpindah *cluster* sehingga menimbulkan perubahan nilai *centroid cluster*.

Untuk memudahkan penjelasan mengenai *K-means* berikut sebuah kasus *data mining cluster*ing, Terdapat data Nilai mahasiswa sebagai Berikut:

**Tabel 1** Tabel Data

| Data Ke | NIM | N. Tugas | *Unit Test* | *Mid test* | Kehadiran | Status Tinggal | Bahasa | Tugas Akhir |
|---|---|---|---|---|---|---|---|---|
| 1 | 1482002 | 100 | 100 | 100 | 96 | 1 | 0 | 0 |
| 2 | 1481007 | 64 | 75 | 67 | 54 | 1 | 0 | 0 |
| 3 | 1382002 | 95 | 75 | 69 | 89 | 0 | 0 | 0 |
| 4 | 1381019 | 96 | 100 | 94 | 96 | 0 | 0 | 0 |
| 5 | 1482017 | 54 | 100 | 77 | 45 | 1 | 0 | 0 |
| 6 | 1382009 | 95 | 50 | 78 | 64 | 0 | 0 | 0 |
| 7 | 1381030 | 94 | 100 | 81 | 71 | 1 | 1 | 1 |
| 8 | 1482014 | 94 | 100 | 92 | 100 | 1 | 1 | 1 |
| 9 | 1482018 | 24 | 25 | 0 | 32 | 0 | 1 | 1 |
| 10 | 1481002 | 96 | 100 | 90 | 100 | 1 | 1 | 1 |

Status tinggal *inside* = 1, *Outside* = 0, Bahasa Indonesia = 0, Bahasa Inggris = 1, tugas akhir Teori = 0, Projek = 1. *Clustering* yang di harapkan mampu menghasilkan :

1. Mahasiswa yang memiliki karakteristik nilai yang sama berada dalam kelompok yang sama.
2. Mahasiswa yang memiliki karakteristik nilai yang berbeda akan di kelompokkan pada kelompok yang lain.

**Langkah 1** Tentukan Jumlah *Cluster* yang di ingin kan (K = 3)

**Langkah 2** Pilih *Centroid* Awal

M1 = Sebagai Mahasiswa yang lulus dengan nilai Bagus (*Centroid* awal adalah data ke 1)
M2 = Sebagai Mahasiswa yang lulus dengan nilai Cukup (*Centroid* awal adalah data ke 5)
M3 = Sebagai Mahasiswa yang tidak lulus (*Centroid* awal adalah data ke 9)

**Langkah 3** Hitung Jarak Dengan *Centroid*

Pada langkah ini setiap data akan ditentukan *centroid* terdekatnya dengan rumus euclidean, dan data tersebut akan ditetapkan sebagai anggota kelompok yang terdekat dengan centroid.

$$D(x,y) = \sqrt{(X_{1x} - X_{1y})^2 + (X_{1x} - X_{1y})^2 + \cdots + (X_{kx} - X_{ky})^2} \quad (1)$$

Dimana:
D(x,y) = Jarak data ke x ke pusat *cluster* y
x = data ke x
y = *centroid* data ke x
X kx = Data ke i pada atribut data ke k
X ky = Titik pusat ke j pada atribut ke k

Untuk menghitung nya adalah sebagai berikut:

1. Data Ke 1 (100, 100, 100, 96, 1, 0, 0)

Jarak Ke – M1

$\sqrt{(100-100)^2+(100-100)^2+(100-100)^2+(96-96)^2+(1-1)^2+(0-0)^2+(0-0)^2}$ = 0

Jarak Ke – M2

$\sqrt{(100-54)^2+(100-100)^2+(100-77)^2+(96-45)^2+(1-1)^2+(0-0)^2+(0-0)^2}$ = 72

Jarak Ke – M3

$\sqrt{(100-24)^2+(100-25)^2+(100-0)^2+(96-32)^2+(1-0)^2+(0-1)^2+(0-1)^2}$ = 160

Berdasarkan Hasil di atas jarak terdekat adalah M1 maka data ke 1 masuk ke dalam *cluster* 1

2. Data Ke 2 (64, 75, 67, 54, 1, 0, 0)

Jarak Ke – M1

$\sqrt{(64-100)^2+(75-100)^2+(67-100)^2+(54-96)^2+(1-1)^2+(0-0)^2+(0-0)^2}$ = 69

Jarak Ke – M2

$\sqrt{(64-54)^2+(75-100)^2+(67-77)^2+(54-45)^2+(1-1)^2+(0-0)^2+(0-0)^2}$ = 30

Jarak Ke – M3

$\sqrt{(64-24)^2+(75-25)^2+(67-0)^2+(54-32)^2+(1-0)^2+(0-1)^2+(0-1)^2}$ = 95

Berdasarkan Hasil di atas jarak terdekat adalah M2 maka data ke 1 masuk ke dalam *cluster* 2

3. Data Ke 3 (95, 75, 69, 89, 0, 0, 0)

Jarak Ke – M1

$\sqrt{(95-100)^2+(75-100)^2+(69-100)^2+(89-96)^2+(0-1)^2+(0-0)^2+(0-0)^2}$ = 41

Jarak Ke – M2

$\sqrt{(95-54)^2+(75-100)^2+(69-77)^2+(89-45)^2+(0-1)^2+(0-0)^2+(0-0)^2}$ = 66

Jarak Ke – M3

$\sqrt{(95-24)^2+(75-25)^2+(69-0)^2+(89-32)^2+(0-0)^2+(0-1)^2+(0-1)^2}$ = 125

Berdasarkan Hasil di atas jarak terdekat adalah M1 maka data ke 1 masuk ke dalam *cluster* 1
Berikut Tabel Setelah dilakukannya perhitungan sampai data ke 10 :

**Tabel 2** Tabel Hasil *Euclidean* Iterasi 1

| N. Tugas | *Unit Test* | *Mid test* | Kehadiran | Status Tinggal | Bahasa | Tugas Akhir | M1 | M2 | M3 | *Cluster* |
|---|---|---|---|---|---|---|---|---|---|---|
| 100 | 100 | 100 | 96 | 1 | 0 | 0 | 0 | 72 | 160 | 1 |
| 64 | 75 | 67 | 54 | 1 | 0 | 0 | 69 | 30 | 95 | 2 |
| 95 | 75 | 69 | 89 | 0 | 0 | 0 | 41 | 66 | 125 | 1 |
| 96 | 100 | 94 | 96 | 0 | 0 | 0 | 7 | 68 | 154 | 1 |
| 54 | 100 | 77 | 45 | 1 | 0 | 0 | 72 | 0 | 112 | 2 |
| 95 | 50 | 78 | 64 | 0 | 0 | 0 | 64 | 67 | 113 | 1 |
| 94 | 100 | 81 | 71 | 1 | 1 | 1 | 32 | 48 | 136 | 1 |
| 94 | 100 | 92 | 100 | 1 | 1 | 1 | 11 | 70 | 154 | 1 |
| 24 | 25 | 0 | 32 | 0 | 1 | 1 | 160 | 112 | 0 | 3 |
| 96 | 100 | 90 | 100 | 1 | 1 | 1 | 12 | 70 | 153 | 1 |

Berdasarkan tabel di atas maka : *Cluster* 1 = data {1,3,4,6,7,8,10}, *Cluster* 2 = data {2,5}, dan *Cluster* 3 = data {9}.

Menghitung BCV (jarak antar *cluster*) : M1(100, 100, 100, 96, 1, 0, 0), M2(54, 100, 77, 45, 1, 0, 0), M3(24, 25, 0, 32, 0, 1, 1)

d(M1,M2) = $\sqrt{(100-54)^2+(100-100)^2+(100-77)^2+(96-45)^2+(1-1)^2+(0-0)^2+(0-0)^2}$ = 72

d(M1,M3) = $\sqrt{(100-24)^2+(100-25)^2+(100-0)^2+(96-32)^2+(1-0)^2+(0-1)^2+(0-1)^2}$ = 160

d(M2,M3) = $\sqrt{(54-24)^2+(100-25)^2+(77-0)^2+(45-32)^2+(1-0)^2+(0-1)^2+(0-1)^2}$ = 112

BCV = d(M1,M2)+ d(M1,M3)+ d(M2,M3) = 344

**Tabel 3** Tabel Jarak Terkecil Iterasi 1

| Data Ke | Jarak Terkecil |
|---|---|
| 1 | 0 |
| 2 | 30 |
| 3 | 41 |
| 4 | 7 |
| 5 | 0 |
| 6 | 64 |
| 7 | 32 |
| 8 | 11 |
| 9 | 0 |
| 10 | 12 |

WCV (jarak antar objek *cluster*) = 02 +302 +412 +72 +02 +642 +322 +112 +02 +122 = 7930

Ratio = BCV/WCV = 0,0434

**Langkah 4** Melakukan Pembaruan *Centroid* dengan menghitung *mean* Nilai setiap *cluster*.

1. C1

**Tabel 4** Centroid baru *Cluster* 1-Iterasi 1

| NO | NIM | Tugas | U.Test | *Mid test* | Kehadiran | Status Tinggal | Bahasa | Tugas Akhir |
|---|---|---|---|---|---|---|---|---|
| 1 | 1482002 | 100 | 100 | 100 | 96 | 1 | 0 | 0 |
| 3 | 1382002 | 95 | 75 | 69 | 89 | 0 | 0 | 0 |
| 4 | 1381019 | 96 | 100 | 94 | 96 | 0 | 0 | 0 |
| 6 | 1382009 | 95 | 50 | 78 | 64 | 0 | 0 | 0 |
| 7 | 1381030 | 94 | 100 | 81 | 71 | 1 | 1 | 1 |
| 8 | 1482014 | 94 | 100 | 92 | 100 | 1 | 1 | 1 |
| 10 | 1481002 | 96 | 100 | 90 | 100 | 1 | 1 | 1 |
| Centroid 1 | | 96 | 89 | 86 | 88 | 1 | 0 | 0 |

2. C2

**Tabel 5** Centroid Baru *Cluster* 2-Iterasi 1

| NO | NIM | Tugas | U.Test | *Mid test* | Kehadiran | Status Tinggal | Bahasa | Tugas Akhir |
|---|---|---|---|---|---|---|---|---|
| 2 | 1481007 | 64 | 75 | 67 | 54 | 1 | 0 | 0 |
| 5 | 1482017 | 54 | 100 | 77 | 45 | 1 | 0 | 0 |
| Centroid 2 | | 59 | 88 | 72 | 50 | 1 | 0 | 0 |

3. C3

**Tabel 6** Centroid Baru *Cluster* 3-Iterasi

| NO | NIM | Tugas | U.Test | *Mid test* | Kehadiran | Status Tinggal | Bahasa | Tugas Akhir |
|---|---|---|---|---|---|---|---|---|
| 9 | 1482018 | 24 | 25 | 0 | 32 | 0 | 1 | 1 |
| Centroid 3 | | 24 | 25 | 0 | 32 | 0 | 1 | 1 |

Maka di dapat nilai Centroid baru dari M1, M2, dan M3.

**Langkah 5** mengulangi langkah 3 dan 4, sampai tidak ada lagi perpindahan anggota pada *cluster* dan tidak ada perubahan rasio *cluster*. Pada iterasi ke 3 tidak ada lagi perubahan ratio proses iterasi berhenti.

Hasil *Cluster* yang di dapat setelah iterasi ke 3 adalah sebagai berikut:

1. C1

**Tabel 7** *Cluster* 1

| NO | NIM | Tugas | U.Test | *Mid test* | Kehadiran | Status Tinggal | Bahasa | Tugas Akhir |
|---|---|---|---|---|---|---|---|---|
| 1 | 1482002 | 100 | 100 | 100 | 96 | 1 | 0 | 0 |
| 3 | 1382002 | 95 | 75 | 69 | 89 | 0 | 0 | 0 |
| 4 | 1381019 | 96 | 100 | 94 | 96 | 0 | 0 | 0 |
| 6 | 1382009 | 95 | 50 | 78 | 64 | 0 | 0 | 0 |
| 7 | 1381030 | 94 | 100 | 81 | 71 | 1 | 1 | 1 |
| 8 | 1482014 | 94 | 100 | 92 | 100 | 1 | 1 | 1 |
| 10 | 1481002 | 96 | 100 | 90 | 100 | 1 | 1 | 1 |

2. C2

**Tabel 8** *Cluster* 2

| NO | NIM | Tugas | U.Test | *Mid test* | Kehadiran | Status Tinggal | Bahasa | Tugas Akhir |
|---|---|---|---|---|---|---|---|---|
| 2 | 1481007 | 64 | 75 | 67 | 54 | 1 | 0 | 0 |
| 5 | 1482017 | 54 | 100 | 77 | 45 | 1 | 0 | 0 |

3. C2

**Tabel 9** *Cluster* 3

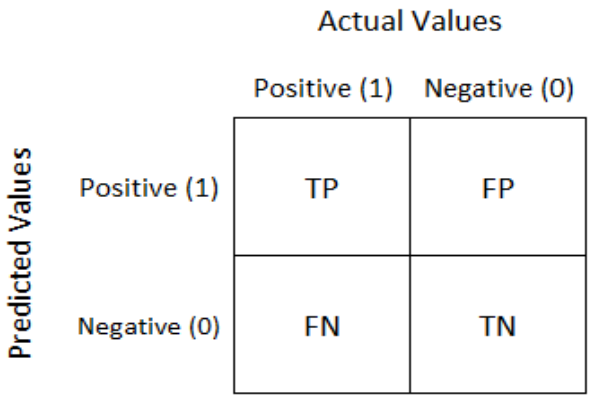
| NO | NIM | Tugas | U.Test | *Mid test* | Kehadiran | Status Tinggal | Bahasa | Tugas Akhir |
|---|---|---|---|---|---|---|---|---|
| 9 | 1482018 | 24 | 25 | 0 | 32 | 0 | 1 | 1 |

Berdasarkan Hasil di atas maka :

1. Terdapat 7 Mahasiswa yang lulus mata kuliah dengan nilai baik.
2. Terdapat 2 mahasiswa yang lulus mata kuliah dengan nilai cukup.
3. Terdapat 1 mahasiswa yang tidak lulus dari mata kuliah tersebut

***Performance Evaluation***

Dalam menghitung *performance evaluation*, peneliti menggunakan metode *confusion matrix* untuk menghitung tingkat akurasi, presisi, dan *recall* hasil prediksi. *Confusion matrix* adalah metrik yang memperlihatkan kebenaran dan kesalahan prediksi data dari hasil sebuah algoritma[13].




**Gambar 2** Confusion Matrix

Berdasarkan hasil prediksi di atas maka akan di hitung akurasi, presisi, dan *recall* dengan menggunakan confussion matrix, berikut perhitungannya:

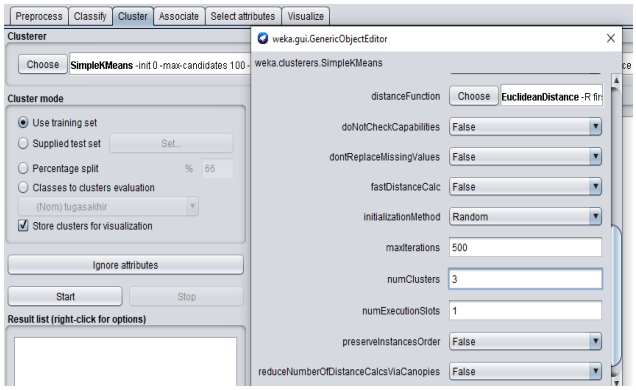
**Tabel 10** Hasil Confusion Matrix

| **Data Prediksi** | | **Data Aktual** | | **Confusion Matrix** | | | |
|---|---|---|---|---|---|---|---|
| *Lulus* | *Tidak* | *Lulus* | *Tidak* | *TP* | *TN* | *FP* | *FN* |
| 9 | 1 | 8 | 2 | 8 | 1 | 1 | 1 |

1. Akurasi = (TP + TN) / (TP + TN + FP + FN)) * 100% = (8+1)/(8+1+1+1)*100% = 81%
2. Precision = (TP / (TP + FP)) * 100% = (8/(8+1)*100% = 88%
3. *Recall* = (TP / (TP + FN)) * 100% = (8/(8+1)*100% = 88%

## 3. Hasil

Setelah dilakukannya *preprocessing* data menggunakan tahapan di atas untuk memperbaiki *dataset* sebelum dilakukannya proses *mining* maka data yang dapat diolah sebanyak 118 data mahasiswa. data tersebut berisi data nilai mahasiswa Fakultas Teknologi Informasi angkatan 2015, 2016, 2017, dan 2018 pada tahun ajaran 2018-2019. Pengumpulan data dilakukan dengan cara meminta secara langsung dengan dosen pengajar mata kuliah, data nilai didapatkan dari 7 mata kuliah. proses mining dilakukan dengan bantuan menggunakan aplikasi WEKA (*Waikato Environment for Knowledge Analysis*).



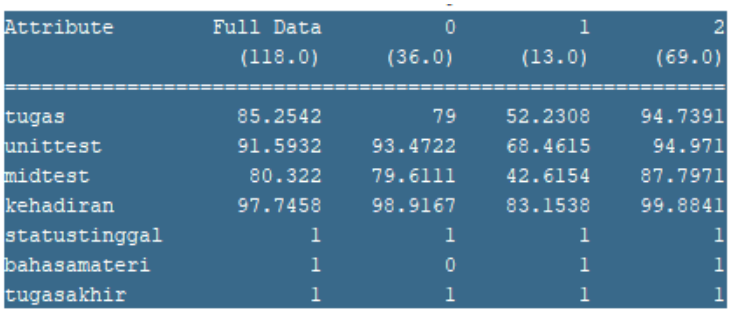

**Gambar 3** Pemilihan Algoritma *K-means* dan Nilai K=3

*Centroid* Awal yang dipilih secara *random*:

**Tabel 11** Centroid Awal

| NIM | Tugas | *Unit Test* | Mid | Kehadiran | Status Tinggal | Bahasa Materi | Tugas Akhir | *Cluster* |
|---|---|---|---|---|---|---|---|---|
| 1681022 | 72 | 100 | 80 | 100 | 1 | 0 | 1 | C0 |
| 1681008 | 86 | 100 | 75.0 | 100 | 1 | 1 | 1 | C1 |
| 1682014 | 74 | 100 | 90 | 100 | 1 | 1 | 1 | C2 |

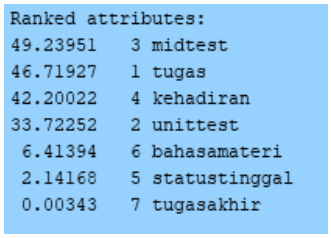
Banyaknya iterasi yang dilakukan adalah sebanyak 6 iterasi, hasil iterasi setelah dilakukannya proses *mining* dengan *K- means* adalah sebagai berikut:



| Attribute | Full Data (118.0) | 0 (36.0) | 1 (13.0) | 2 (69.0) |
|---|---|---|---|---|
| tugas | 85.2542 | 79 | 52.2308 | 94.7391 |
| unittest | 91.5932 | 93.4722 | 68.4615 | 94.971 |
| midtest | 80.322 | 79.6111 | 42.6154 | 87.7971 |
| kehadiran | 97.7458 | 98.9167 | 83.1538 | 99.8841 |
| statustinggal | 1 | 1 | 1 | 1 |
| bahasamateri | 1 | 0 | 1 | 1 |
| tugasakhir | 1 | 1 | 1 | 1 |

**Gambar 4** Hasil Iterasi

Berdasarkan data di atas pada k1 terdapat 13 mahasiswa yang di prediksi tidak lulus karena memiliki nilai tengah rendah, pada k2 terdapat 69 mahasiswa yang diprediksi lulus dengan nilai cukup karna memiliki nilai tengah tinggi, dan pada k0 terdapat 36 mahasiswa yang di prediksi lulus dengan nilai cukup. Setelah selesai dilakukannya prediksi kelulusan mahasiswa dalam mata kuliah, peneliti melakukan pemilihan atribut yang mempengaruhi kelulusan mata kuliah menggunakan *chi square atribute* selection untuk mengetahui hal-hal apa saja yang dapat mempengaruhi mahasiswa lulus di sebuah mata kuliah dengan menggunakan aplikasi WEKA, maka hasil yang di dapat adalah sebagai berikut:



```
Ranked attributes:
49.23951    3 midtest
46.71927    1 tugas
42.20022    4 kehadiran
33.72252    2 unittest
 6.41394    6 bahasamateri
 2.14168    5 statustinggal
 0.00343    7 tugasakhir
```

**Gambar 5** Hasil *Chi square atribute* Selection

Berdasarkan hasil *chi square atribute* selection di atas, maka, atribut yang mempengaruhi hasil prediksi dari yang tertinggi hingga terendah:

1. Nilai Tugas, mempengaruhi prediksi kelulusan terhadap 46% dari 118 mahasiswa.
2. Nilai *Mid test* mempengaruhi prediksi kelulusan terhadap 49% dari 118 mahasiswa.
3. Tugas Akhir tidak terlalu mempengaruhi prediksi kelulusan tersebut.

Setelah dilakukannya prediksi peneliti menggunakan *confusion matrix* untuk menghitung tingkat akurasi, presisi, dan *recall* dari hasil prediksi, jumlah mahasiswa yang sebenarnya lulus sebanyak 109 mahasiswa dan yang tidak lulus adalah sebanyak 9 mahasiswa, dan hasil prediksi menunjukkan jumlah mahasiswa yang lulus sebanyak 105 mahasiswa dan yang tidak lulus sebanyak 13 mahasiswa. Tingkat akurasi, presisi, dan *recall* menggunakan *confusion matrix* adalah sebagai berikut:

**Tabel 12** Confusion Matrix

| **Data Prediksi** | | **Data Aktual** | | **Confusion Matrix** | | | |
|---|---|---|---|---|---|---|---|
| *Lulus* | *Tidak* | *Lulus* | *Tidak* | *TP* | *TN* | *FP* | *FN* |
| 105 | 17 | 109 | 9 | 105 | 4 | 4 | 9 |

**Tabel 13** Hasil Akurasi, Presisi dan Recall

| **Akurasi** | **Presisi** | **Recall** |
|---|---|---|
| 93% | 96% | 92% |

## 4. Pembahasan/Kesimpulan

Berdasarkan hasil yang didapatkan dari perhitungan di atas maka di dapatkan informasi sebagai berikut:

1. Centroid akhir yang di dapat dari proses mining

**Tabel 14** Centroid Akhir

| **Tugas** | ***Unit Test*** | **Mid** | **Kehadiran** | **S.Tinggal** | **Bahasa** | **Tugas** | ***Cluster*** |
|---|---|---|---|---|---|---|---|
| 79 | 93 | 79 | 98 | 1 | 0 | 1 | C0 |
| 52 | 68 | 42 | 83 | 1 | 1 | 1 | C1 |
| 94 | 94 | 87 | 99 | 1 | 1 | 1 | C2 |

2. Terdapat 13 mahasiswa pada *cluster* 1 dengan rata-rata nilai rendah, sehingga pada *cluster* 1 merupakan mahasiswa yang tidak lulus.
3. Terdapat 36 mahasiswa pada *cluster* 0 dengan rata-rata nilai cukup, sehingga pada *cluster* 0 merupakan mahasiswa yang lulus dengan nilai cukup.
4. Terdapat 69 mahasiswa pada *cluster* 2 dengan rata-rata nilai tinggi, sehingga pada *cluster* 2 merupakan mahasiswa yang lulus dengan nilai tinggi.
5. Bahasa materi yang menggunakan Bahasa inggris, mempengaruhi mahasiswa yang terdapat pada *cluster* 0 yaitu mahasiswa yang lulus dengan nilai cukup

**Kesimpulan**

Berdasarkan hasil yang di dapat, kesimpulan yang di dapat pada penelitian ini adalah:

1. Dengan menggunakan algoritma *K-means* untuk memprediksi kelulusan mahasiswa, hasil prediksi sebanyak 105 mahasiswa yang diprediksi lulus, dan 13 mahasiswa diprediksi tidak lulus.
2. Dari hasil iterasi, materi yang menggunakan Bahasa inggris, mempengaruhi mahasiswa yang terdapat pada *cluster* 0 yaitu mahasiswa yang lulus dengan nilai cukup sehingga pemahaman akan Bahasa materi sangat penting bagi mahasiswa untuk meningkatkan kemampuan internalnya. Dan dari hasil *chi square atribute selection* terhadap hasil prediksi, dari atribut yang digunakan yang mempengaruhi prediksi kelulusan mata kuliah adalah Nilai *Mid test*, Nilai Tugas,

dan Kehadiran. Sedangkan atribut yang tidak terlalu mempengaruhi prediksi kelulusan mata kuliah adalah tugas akhir yang diberikan.

3. Hasil perhitungan menggunakan *confusion matrix* menunjukkan bahwa tingkat akurasi prediksi terhadap data aktual adalah 93% yang artinya algoritma K-mean cukup baik untuk memprediksi kelulusan mata kuliah.

**Diskusi/Saran**

Saran yang diberikan oleh peneliti untuk pengembangan penelitian yang akan dilakukan kemudian hari adalah :

1. Menggunakan algoritma *data mining* yang lain seperti *Fuzzy C means* dan *K-medoid*, guna membandingkan akurasi dalam prediksi kelulusan mata kuliah.
2. Menggunakan atribut-atribut tambahan yang belum digunakan pada penelitian ini seperti mahasiswa mengikuti organisasi atau tidak, jenis kelamin mahasiswa, latar belakang pendidikan, perilaku dosen mengajar, dan lain sebagainya, untuk pengembangan penelitian selanjutnya.
3. Menggunakan *dataset* yang lebih banyak lagi, agar dapat memberikan potensi yang lebih baik lagi terhadap mata kuliah yang lain karna pada penelitian ini menggunakan mata kuliah yang memiliki ikatan dengan mata kuliah lain.
4. Melakukan perancangan dan implementasi dalam bentuk aplikasi, guna lebih memudahkan mahasiswa untuk melakukan antisipasi diri agar tidak terjadi keterlambatan mengambil mata kuliah yang lain sehingga dapat membuat mahasiswa lulus tidak tepat waktu.

## 5. Referensi


[1] Agusta, Yudhi. 2007. '*K-means* penerapan permasalahan dan metode terkait'. Jurnal Sistem dan Informatika, Vol 3.

[2] Baradwaj, B. K., & Pal, S. (2011). Mining Educational Data to Analyze Students" Performance. International Journal of Advanced Computer Science and Applications.

[3] Borkar, S., & Rajeswari, K. (2014). Atributes Selection for Predicting Student's Academic Performance using Education *Data mining* and Artificial Neural Network. International Journal of Computer Applications .

[4] C, D. A., Baskoro, D. A., Ambarwati, L., & Wicaksana , I. S. (2013). Belajar *Data mining* Dengan Rapid Miner. Jakarta: Remi Sanjaya.

[5] Darmi, Y., & Setiawan, A. (2016). PENERAPAN METODE *CLUSTERING K-MEANS* DALAM PENGELOMPOKAN PENJUALAN PRODUK. Jurnal Media Infotama, 148.

[6] Herdianto. (2013). Prediksi Kerusakan Motor Induksi Menggunakan Metode Jaringan Saraf Tiruan Backpropagation ". Universitas Sumatera Utara, 8.

[7] Indrawan, B. R. (2018). Penerapan Algoritma *K-means* Untuk Menentukan Strategi Promosi Universitas Islam Negeri Sunan Kalijaga Yogyakarta. 34.

[8] Jaya, D. I. (2006). Perancangan dan pembuatan perangkat lunak berbasis jaringan syaraf tiruan untuk memprediksi harga saham dengan menggunakan metode backpropagation. Scientific Repository.

[9] Murti, D. H., Suciati, N., & Nanjaya, D. J. (2005). *CLUSTERING* DATA NON-NUMERIK DENGAN PENDEKATAN ALGORITMA *K-MEANS* DAN HAMMING DISTANCE STUDI KASUS BIRO JODOH . JUTI, 48.

[10] Praja, B. S., Kusuma, D. D., & Setianingsih, C. (2019). APLICATION OF *K-MEANS CLUSTERING* METHOD IN PASSENGER AND SHIP TRANSPORT DATA GROUPING IN INDONESIA. e-Proceeding of Engineering, 1442.

[11] Purnamasari, D., Henharta, J., Sasmita, Y. P., Ihsani, F., & Wicaksana, I. S. (2013). GET EASY USING WEKA. Jakarta: Dapur Buku.

[12] Sibarani, R., & Chafid. (2018). ALGORITHMA *K-MEANS CLUSTERING* STRATEGI PEMASARAN PENERIMAAN MAHASISSWA BARU UNIVERSITAS SATYA NEGARA INDONDESIA [ALGORITHMA *K-*

*MEANS CLUSTERING* STRATEGY MARKETING ADMISSION UNIVERSITAS SATYA NEGARA INDONESIA]. Seminar Nasional Cendekiawan , 687.

[13] S, R. (2015). Python machine learning. Packt Publishing Ltd, 190.

[14] Subkhan, M. (2010). Algoritma *Cluster*ing.

[15] Sugiono, Nurdian, S., Linawati, S., Safitri, R. A., & Saputra, E. P. (2019). Pengelompokan Perilaku Mahasiswa Pada Perkuliahan E-Learning dengan *K-means Clustering* . Jurnal Kajian Ilmiah Universitas Bhayangkara Jakarta Raya , 128.

[16] Suprayogi. (2018). *Data mining Cluster*ing.