ANALISIS PERBANDINGAN ALGORITMA NAIVE BAYES DAN C.45 DALAM KLASIFIKASI DATA MINING UNTUK MEMPREDIKSI KELULUSAN

Shelly Janu Setyaning Tyas, Mita Febianah[1], Farkhatus Solikhah[2], Amelia Luthfi Kamil[3], Wildan Aprizal Arifin[5]
Sistem Informasi Kelautan[1,2,3,4,5]
Universitas Pendidikan Indonesia[1,2,3,4,5]
e-mail: shellyjanu@upi.edu, mitafebianah@upi.edu, farkhatus.solikhah@upi.edu, amelialuthfi@upi.edu, willdanarifin@upi.edu

Abstrak : Tingkat akurasi data dalam kehidupan sehari-hari sangatlah dibutuhkan karena melihat perkembangan teknologi informasi yang semakin maju. Analisa pengelolaan data menjadi informasi yang bisa memberikan pengetahuan yaitu caranya dengan menggunakan sistem *data mining*. Algoritma yang sering digunakan untuk memprediksi kelulusan yaitu *Naive Bayes dan C.45*. Tujuan dari penelitian ini yaitu untuk membandingkan *algoritma Naive Bayes* dan *algoritma C.45* dalam hal keakuratan prediksi kelulusan. Metode yang digunakan yaitu dengan studi literatur dari berbagai sumber terkait serta memahami data-data yang ada pada sumber yang berhubungan dengan topik metode klasifikasi algoritma *Naive Bayes* dan *C.45* dalam sebuah sistem *data mining*. Hasil dari penelitian ini menunjukan bahwa pengklasifikasian dengan menggunakan algoritma *Naive Bayes* tingkat keakuratannya lebih tinggi dibandingkan algoritma *C.45*.
Kata Kunci : *C.45*, data mining, keakuratan, *naive bayes*

1. Pendahuluan

Tingkat akurasi data dalam kehidupan sehari-hari sangatlah dibutuhkan karena melihat perkembangan teknologi informasi yang semakin maju. Menurut Huda (2010) dalam penentuan tiap keputusan dalam kondisi tertentu hal yang perlu diperhatikan yaitu informasinya, maka dari itu tersedianya informasi telah menjadi media untuk menganalisa dan merangkum pengetahuan dari data yang berguna dalam pengambilan keputusan. Namun dalam mengambil keputusan pengetahuan dari data pada suatu informasi saja tidak cukup. Dibutuhkan juga sebuah analisa untuk menghasilkan bahan pertimbangan dari informasi yang telah disediakan. Analisa pengelolaan data menjadi informasi yang bisa memberikan pengetahuan yaitu caranya dengan menggunakan sistem *data mining* (Ali, 2013).

Pada sebuah kasus dapat dilihat dari kecenderungan dari segi tatanan ataupun perkiraanya di waktu mendatang yaitu dengan menggunakan *data mining*. Penerapan tahapan dan teknik pada *data mining* dalam kehidupan nyata ada beragam, salah satunya yaitu dengan teknik klasifikasi. Menurut Kurniawan (2018) klasifikasi ini merupakan bentuk dasar dari analisis data, sedangkan menurut Bansal, *et al* (2017) klasifikasi yaitu teknik yang digunakan untuk keanggotaan kelompok menurut data-data yang sudah tersedia. Dapat dikatakan juga klasifikasi ini merupakan salah satu problema mendasar dan tugas utama pada *data mining* (Zhang dan Harry, 2004 dalam Syarli, 2016). Klasifikasi dapat digunakan untuk beragam kasus sebagai pembentukan aturan terhadap sebuah data.

Sebelumnya telah banyak penelitian untuk memprediksi kelulusan dengan metode klasifikasi, salah satunya yaitu dengan *algoritma Naive Bayes* dan *algoritma C.45*. Menurut Kurniawan (2018) *algoritma Naive Bayes* dan *C.45* ini sendiri saat ini memang sedang popular hal ini karena tingkat akurasi dari kedua algoritma ini tinggi. Maka dari itu penelitian ini bertujuan untuk membandingkan tingkat keakuratan dari *algoritma Naive Bayes* dan *algoritma C.45* dalam hal keakuratan prediksi kelulusan.

2. Kajian Pustaka

Data Mining

Data mining merupakan pencarian pada pola atau aturan dari sebuah data yang kapasitasnya sangat besar untuk menemukan informasi baru. (Davies dan Beynon, 2004). *Data mining* juga dapat disebut sebagai proses yang digunakan untuk menentukan sebuah struktur data (Roiger, 2017). Karakteristik dari *data mining* (Davies dan Beynon) yaitu : (i) berkaitan dengan penemuan sesuatu yang baru yang tidak diketahui sebelumnya terkait pola data tertentu; (ii) data yang digunakan merupakan data yang besar untuk membuat hasil yang besar; (iii) dapat digunakan untuk penentuan strategi dan keputusan yang kritis.

Algoritma Naive Bayes

Naive Bayes merupakan sebuah pengklasifikasian atau penggolongan data yang menghitung kemungkinan dari dataset yang tersedia (Saleh, 2015). Sedangkan menurut Bustami (2013) *Naive Bayes* merupakan pengklasifikasian untuk memprediksi peluang masa depan dengan metode probabilitas dan statistik sesuai dengan pengalaman di waktu sebelumnya.

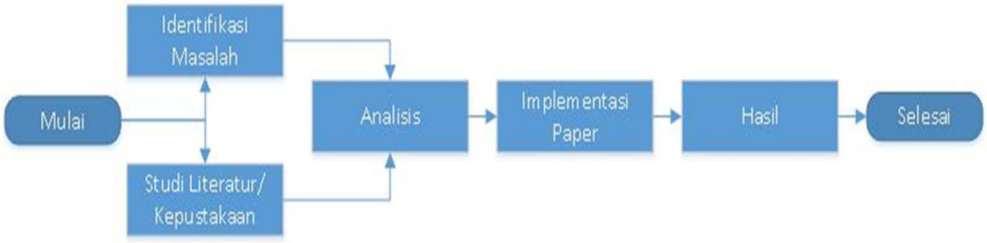
Kelebihan dari *algoritma Naive Bayes* adalah data yang di butuhkan untuk menetapkan perkiraan parameter dalam proses penggolongan dalam menggunakan metode ini hanya memerlukan jumlah data pelatihan yang kecil (Saleh, n.d). Sedangkan menurut Syarli (2016) kelebihan *Naive Bayes* yaitu mudah diimplementasikan dan pada banyak kasus memberikan hasil yang baik, kemudian kekurangnnya yaitu tidak terkaitnya antar fitur atau bersifat *independent*, sedangkan pada kenyataannya keterkaitan itu harus ada dan tidak dapat dimodelkan oleh Naive *Bayesian Classifier.*

Algoritma C.45

Algoritma C4.5 adalah salah satu diantara banyak algoritma yang bisa dipergunakan untuk membuat pohon keputusan (Mardi, n.d). Sedangkan menurut Elisa (2017) penggunaan *algoritma C.45* untuk mengklasfikasi data yang memiliki atribut-atribut kuantitatif (numerik) dan kategorial. Hasil dari proses klasifikasi ini berupa aturan-aturan yang dipergunakan untuk memperkirakan nilai atribut bersifat diskret atau kualitatif dari rekaman yang baru. *Algortima C4.5* merupakan pengembangan dari algortima ID3 yang dapat menangani masalah informasi yang tidak tersedia, serta dapat menangani data *kontinu* dan *pruning*.

3. Metode Penelitian

Pada penelitian ini dilakukan secara sistematik sehingga pada proses penelitian terdapat pedoman agar penelitian dapat mencapai hasil yang diharapkan dan menghindari penyimpangan dari tujuan penelitian. Adapun tahapan penelitian adalah sebagai berikut:



Gambar 1. Metode Penelitian

Tahapan dari metodologi penelitian dijelaskan sebagai berikut :

a. Identifikasi Masalah
   Tahapan ini dilakukan yaitu untuk melakukan pendefinisian terhadap masalah penelitian dengan menguraikan poin penting permasalahan sehingga penulis mendapatkan landasan untuk menguraikan masalah penelitian.
b. Studi Literatur
   Penelitian dilakukan secara literatur dari berbagai sumber terkait serta memahami data-data yang ada pada sumber yang berhubungan dengan topik metode klasifikasi *algoritma Naive Bayes* dan *C.45* dalam sebuah sistem data mining. Dilakukan juga pengkajian tentang teori yang berhubungan dengan topik penelitian yang telah dilakukan pada penelitian terdahulu serta perkembangan berbagai teori saat ini.
c. Analisis
   Pada tahap ini dilakukan proses pengkajian, penguraian dan pemecahan terhadap suatu masalah dan tujuan penelitian.
d. Implementasi Paper
   Pada tahap ini dilakukan penerapan hasil analisis kedalam sebuah paper.

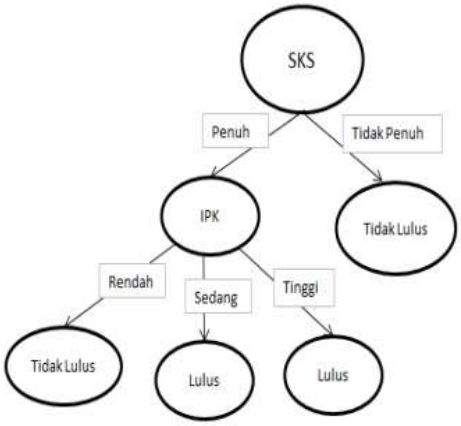
4. Hasil dan Pembahasan

Pada penelitian ini analisa dengan melakukan perbandingan dua metode yaitu *algoritma C.45* dan metode *algoritma Naive Bayes* dalam sebuah sistem *data mining*. Disini kita membandingkan kedua metode dengan menggunakan data dari dua sumber yang mengambil tema mengenai prediksi kelulusan mahasiswa perguruan tinggi. Sumber pertama pada data mahasiswa dibawah ini menggunakan perhitungan *algoritma C.45*. Proses pada pengolahan data ini menggunakan metode *C.45* berfungsi membangun sebuah pohon keputusan terdapat beberapa proses yaitu:

1. Menghitung hasil penjumlahan data, penjumlahan data ini berdasarkan banyaknya atribut hasil dengan memenuhi syarat yang sudah ditetapkan.
2. Menetapkan atribut tersebut dan gunakan untuk *Node*. *Node* yaitu salah satu atribut dimana nilai gainnya tertinggi dari atribut lain.
3. Membuat percabangan untuk semua anggota dari *Node.*
4. Memeriksa bila ada anggota *Node* yang bernilai nol, namun jika hasil ada yang bernilai nol maka tentukanlah mana yang cocok menjafi daun dari pohon keputusan. Lakukan sampai keseluruhan nilai entropy anggota dari *Node* bernilai nol itu artinya proses berhenti.

5. Jika muncul nilai entropy melebihi nol yang asalnya dari saah satu anggota dari *Node*, maka lakukan ulang proses sebelumnya dari awal hingga semua *Node* bernilai semua nol.

Setelah melalui beberapa proses perhitungan menggunakan metode pohon keputusan dari data kemudian didapatkan hasil dari nilai grain tertinggi yaitu pada SKS. Maka SKS disini ditempatkan menjadi sebuah akar dari pohon keputusan. Dan IPK menjadi faktor penentuan kelulusan baru. Bisa dilihat lebih jelas pada gambar berikut:

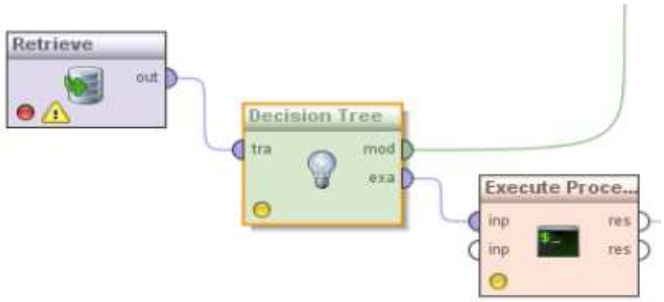


(Sumber: Prayoga, n.d)

Gambar 2. Pohon Keputusan Akhir

Pada Gambar 2 dijelaskan bahwa terdapat 2 macam anggota yaitu anggota penuh (lulus) dan anggota tidak penuh (tidak lulus). Dimana IPK terdapat 3 anggota yaitu rendah (tidak lulus), sedang (lulus), tinggi (lulus). Dan mengapa pohon keputusan ini hanya sampai pada IPK, hal ini dikarenakan nilai tersebut berada antara Anggota lulus dan tidak lulus terdapat nilai 0, maka keputusan nya bisa langsung didapatkan. Kemudian juga terlihat etika dan prestasi tidak mempengaruhi kelulusan mahasiswa tepat waktu.

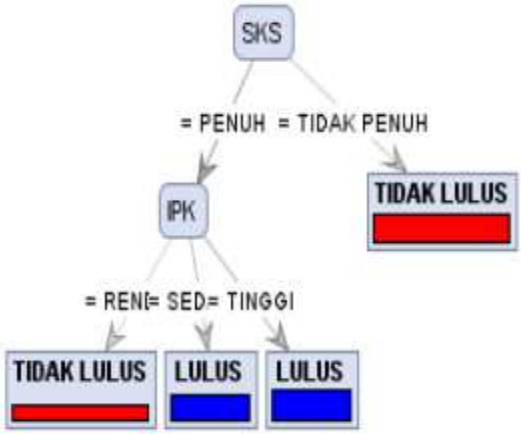
Pengujian selanjutnya yaitu terhadap data sampel menggunakan tools yang ada pada aplikasi exel yaitu tools *Rapidminer* 5 dimulai dari proses koneksi antara basis data sampel ke operator dan selanjutnya validasi seperti pada Gambar 3 berikut:



(Sumber: Prayoga, n.d)

Gambar 3. Koneksi Rapidminer 5

Dari proses koneksi Rapidminer 5 diaatas mendapatkan hasil yang sama seperti dengan proses perhitungan secara manual (Gambar 1.) hingga mendapatkan hasil pohon keputusan seperti berikut ini :

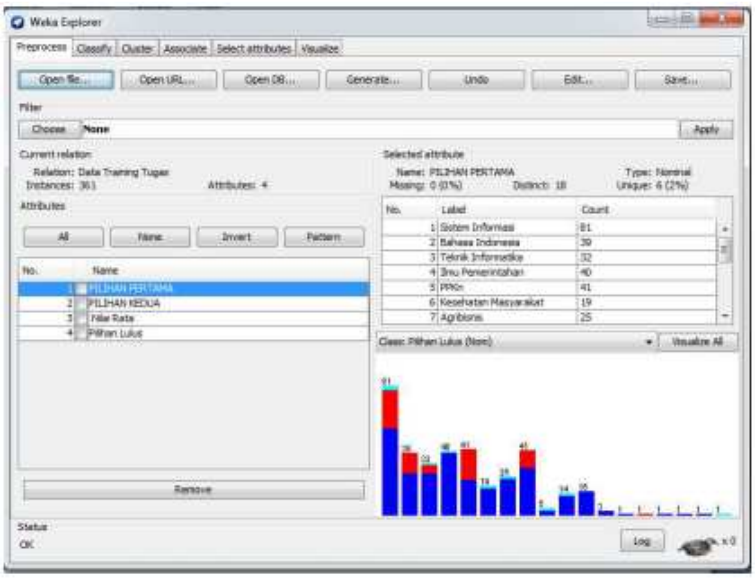


(Sumber: Prayogan, n.d)

Gambar 4. Pohon Keputusan Pada Tools Rapidmider 5

Dalam penggunaan algoritma C.45 mendapatkan hasil pengukuran akurasi dalam memprediksi kelulusan tepat waktu yaitu sebesar 92,60% +\- 1.60%

Pada studi kasus yang kedua ini menerapkan metode *Naive Bayes* dengan pengolahan data masih pada tema prediksi kelulusan mahasiswa baru perguruan tinggi. Pada metode ini terdapat prose pengujian aplikasi menggunakan WEKA. Dengan aplikasi WEKA ini akan dikelompokkan menurut atribut yang terpilih yaitu ada atribut Prodi, atribut pilihan pertama, atribut pilihan kedua, dan yang terakhir ada atribut nilai rata-rata mahasiswa baru.

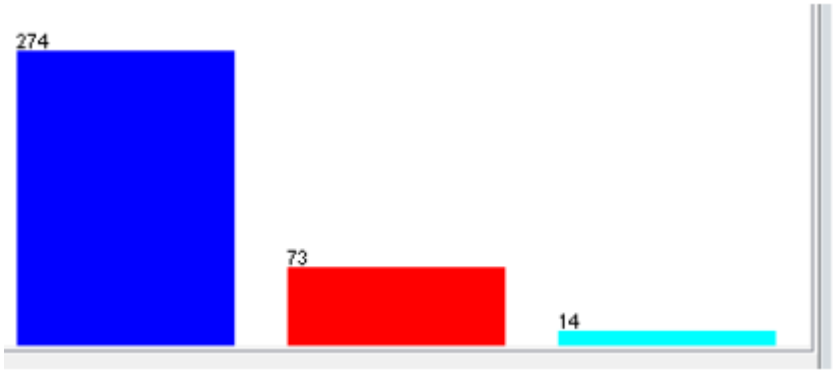


(Sumber: Syarli, 2016)

Gambar 5. Proses Weka Tools

Data tersebut akan memalui proses yang namanya proses (CNBU) *Classifier Naïve Bayes Updateable*. Teknik *classifier* panel dapat melakukan konfigurasi serta menjalankan klasifikasi dimana telah dipilih buat nantinya digunakan di data set. Pada klasifikasi ini sangat dipengaruhi dari atribut lain yang mendorong pengelompokan

penentuan lolos pada pilihan pertama dan pada pilihan kedua, yang nantinya hasil dari klasifikasi tersebut akan disajikan melalui diagram. Setelah memalui proses klasifikasi menunjukkan bahwa presentase pada pilihan pertama lebih unggul mencapai 274 jumlah data sedangkan untuk pilihan kedua yaitu 73 jumlah data dan terakhir pada pilihan tidak lulus mendapat 14 jumlah data.



(Sumber: Syarli, 2016)

Gambar 6. Diagram Hasil Klasifikasi Pilihan Lulusan

Tahapan selanjutnya yaitu tahap evaluasi menunjukkan bahwa data yang telas diklasifikasi dengan benar dan telah sesuai dengan pengelompokan sebanyak 338 data, sedangkan hasil klasifikasi namun tidak sesuai pengelompokan sebanyak 23 data. Pengolahan data menggunakan *confusion matrix* dengan data yang dipakai sebanyak 361 record.

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 271   2   1 |   a = PIL 1
  16  57   0 |   b = PIL 2
   3   1  10 |   c = TIDAK LULUS
```

(Sumber: Syarli, 2016)

Gambar 7. *Confusion Matrix*

Dengan menggunakan 361 record dan hasil dari *confusion matrix* terlihat bahwa sebanyak 271 jumlah record pada kelas a ternyata diprediksi tepat sebagai anggota kelas a sedangkan 3 jumlah record lainnya diprediksi tidak masuk kedalam kelompok anggota pilihan 1. Selanjutnya 57 jumlah record di kelas b ternyata diprediksi tepat sebagai anggota kelas b sedangkan 16 jumlah record lainnya diprediksi tidak tepat masuk ke anggota kelompok data kelas pada pilihan 2. Dan 10 jumlah record lainnya pada kelas c ternyata diprediksi tepat untuk anggota kelas c sedangkan 4 jumalah record diprediksi tidak cocok untuk anggota kelompok data kelas pada pilihan 1 dan 1 jumlah record terakhir ternyata diprediksi bisa masuk anggota kelompok kelas pada pilihan pertama dan juga pilihan kedua.

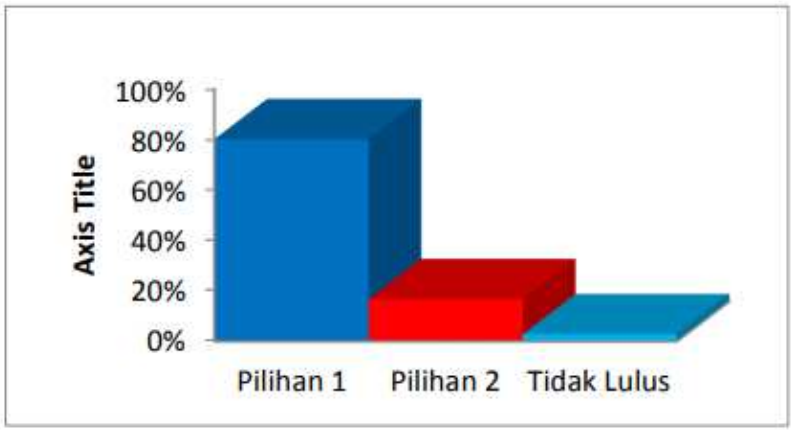
Untuk menentukan nilai presentase keakuratan data menggunakan formula berikut (Syarli, 2016):

Presentase Keakuratan = Jumlah Prediksi Benar/Total Prediksi X 100%

$$\text{Presentase Akurasi} = \frac{271 + 57 + 10}{271 + 2 + 1 + 16 + 57 + 0 + 3 + 1 + 10} \text{ X } 100\%$$

$$= 94\%$$

Nilai keakuratan data ini untuk menunjukkan keefektifan dataset yang sedang diolah yang diterapkan pada metode *Naive Bayes Classification* mencapai 94%.



(Sumber, Syarli, 2016)

Gambar 8. Grafik Hasil Klasifikasi Berdasarkan Presentase Keakuratan

5. Kesimpulan

Dari hasil penjelasan pada uraian diatas dapat disimpulkan bahwa penggunaan metode *algoritma Naive Bayes* nilai keakuratan data untuk menunjukkan keefektifan dataset yang sedang diolah yang diterapkan mencapai 94%. Sedangkan pada *algoritma C.45* mendapatkan hasil pengukuran akurasi dalam memprediksi kelulusan tepat waktu yaitu sebesar 92,60% +\- 1.60%. Maka hal ini memunjukan bahwa untuk memprediksi kelulusan *algoritma Naive Bayes* memiliki klasifikasi tingkat keakuratan yang lebih tinggi dibandingkan *algoritma C.45.*

Daftar Pustaka

Ali, F. (2013). Penerapan Data Mining Untuk Mengetahui Tingkat Kekuatan Beton Yang Dihasilkan Dengan Metode Estimasi Menggunakan Algoritma Linear Regression (Skripsi). Fakultas Ilmu Komputer Universitas Dian Nuswantoro.

Bansal, A., Sharma, M., Goel, S. (2017). Improved K-Mean Clustering Algorithm For Prediction Analysis Using Classification Technique In Data Mining. *International Journal of Computer Applications*, 157(6), 33-40.

Bustami. (2013). Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, TECHSI. *Jurnal Penelitian Teknik Informatika*, 3(2), 127-146.

Davies., Beynon, P. (2004). Database Systems Third Edition. Plgrave Macmillan, New York.

Elisa, E. (2017). Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti. *Jurnal Online Informatika*, 2(1), 36-41.

Huda, N. M. (2010). Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Skripsi). Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Diponegoro.

Kurniawan, Y. I. (2018). Perbandingan Algoritma Naive Bayes dan C.45 Dalam Klasifikasi Data Mining. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 5(4), 455-464.

Mardi, Y. (n.d). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Jurnal Edik Informatika*, 2(2), 213-219.

Prayoga. (n.d). Penerapan Algoritma C.45 Dalam Memprediksi Kelulusan Tepat Waktu Pada Pergruan Tinggi.

Roiger, R. J. (2017). Data Mining: A Tutorial-Based Primer. CRC Press.

Saleh, A. (2015). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Citec Journal*, 2(3), 207-217.

Saleh, A. (n.d). Klasifikasi Metode Naive Bayes Dalam Data Mining Untuk Menentukan Konsentrasi Siswa. Konferensi Nasional Pengembangan Teknologi Informasi dan Komunikasi.

Syarli. (2016). Metode Naive Bayes Untuk Prediksi Kelulusan. *Jurnal Ilmiah Ilmu Komputer*, 2(1), 22-26.