PENGOLAHAN DATA SEDERHANA MENGGUNAKAN R STUDIO

Hesmi Aria Yanti[1]
hesmiaria11@gmail.com

Universitas Muhammadiyah Kotabumi

Abstract: *Simple data processing using the r studio programming language is a data processing process that needs to be done to convert raw data into information. The data processing method uses data acquisition, data input, data selection, division of study programs. split data using k-fold cross-validation, text word cloud model, and model evaluation. Secondary data was acquired in excel format in May 2021, the number of datasets is 71 records with 5 variables, namely number, name, gender, faculty, and study program (Prodi). Data selection aims to select variables that are needed and variables that are not needed are deleted, so that the results of data selection have 2 variables, namely Name and Study Program and a dataset of 71 records. K-fold cross-validation has training data 54 records and testing data 17. The text mining model is visualized with word cloud data, the results of the word cloud testing data test show that there are 4 most important words, including "STI" with a frequency value of 5, "Teacher" a frequency value of 4 , "Law" frequency value 3, and "Agriculture" frequency value 2.*

Keywords: *Data, Processing, R studio*

Abstrak: Pengolahan data sederhana menggunakan bahasa pemprograman r studio merupakan proses pengolahan data yang perlu dilakukan untuk mengubah data mentah menjadi informasi. Metode pengolahan data menggunakan akuisisi data, *input* data, seleksi data, pembagian program studi. split data menggunakan *k-fold cross-validation*, model *text word cloud*, dan evaluasi model. Data sekunder diakuisisi dalam format excel pada bulan mei tahun 2021, jumlah *dataset* yaitu sebanyak 71 *record* dengan 5 *variabel* yaitu nomor, nama, jenis kelamin, fakultas, dan program studi (Prodi). Seleksi data bertujuan memilih variabel yang dibutuhkan dan variabel yang tidak diperlukan dihapus, sehingga hasil seleksi data terdapat 2 variabel yaitu Nama dan Prodi dan *dataset* 71 *record*. *K-fold cross-validation* memiliki *training data* 54 *record* dan *testing data* 17. Model *text mining* divisualisasikan data *word cloaud*, hasil uji *word cloud testing data* menunjukan bahwa terdapat 4 kata terpenting, diantaranya "STI" nilai frekuensi 5, "Keguruan" nilai frekuesi 4, "Hukum" nilai frekuensi 3, dan "Pertanian" nilai frekuensi 2.

Kata Kunci: Data, Pengolahan, R studio

I. PENDAHULUAN

Perkembangan teknologi pada pengelolaan data menggunakan bahasa pemprograman R studio, saat ini telah digunakan berbagai kalangan untuk keperluan analisa data pengolahan data sederhana maupun big data. Pengolahan

[1]*Dosen FTIK Universitas Muhammadiyah Kotabumi*

data sederhana (simple) maupun yang terstruktur, termasuk di bahasa pemrograman R studio. Proses pengolahan data perlu dilakukan untuk mengubah data mentah menjadi informasi, merupakan proses pengumpulan, manipulasi, visualisasi dan evaluasi. Untuk menyederhanakan data mengguakan metode Statistika Deskriptif agar mudah dipahami. Adapun visuaisasi model dalam bentuk tabel, baik tabel frekuensi maupun tabel silang atau dalam bentuk diagram dan grafik seperti diagram batang, kurva dan lainnya (Setiawan 2005).

Menurut Sugiyono (2010), triangulasi sumber berarti peneliti menggunakan teknik pengumpulan data yang sama untuk mendapatkan data dari sumber yang berbeda. Teknik pengumpulan data dapat dilakukan secara langsung dari sumber atau disebut data primer melalui wawancara, kusioner dan data skunder merupakan sumber data yang tersedia atau dikumpulkan dari pihak lain diluar instansi peneliti. R adalah bahasa pemrograman open source yang berhubungan dengan komputasi dan pengolahan data untuk Statistika dan yang berhubungan dengan penampilan grafik menggunakan tools yang disediakan oleh paket-paketnya yang sangat berguna di dalam penelitian dan industri. Versi awal dari R dibuat pada tahun 1992 di Universitas Auckland, New Zealand oleh Ross Ihaka dan Robert Gentleman (Permana et al.2018).

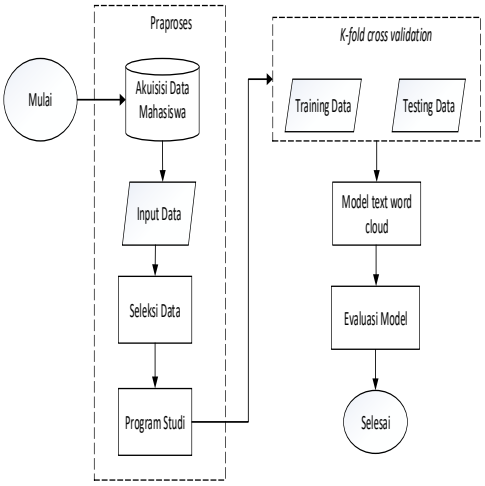
Penggunaan aplikasi r studio pengolahan data linear dan nonlinear untuk identifikasi model, uji statistiks, klasifikasi, analisis dan visualisasi. Kelebihan R lainnya yaitu plot grafik berkualitas penampilan simulasi dapat berupa plot diagram batang, grafik, kurva,wordcloud dan lainnya, termasuk simbol matematika dan rumus jika diperlukan (Budiarto w dan Rachmawati RN. 2013).

Berdasarkan uraian di atas, maka dilakukan penelitian mengenai pengolahan data sederhana pada pemintan mahasiswa menggunakan r studio.

Tujuan dari penelitian ini ialah membuat visualisasi katagori pemintan mahasiswa pada program studi yang ada di Universitas Muhammadiyah kotabumi dengan hasil analisis menggunkan visualisasi dalam bentuk histogram dan *word cloud*.

II. METODE

Penelitian ini terdiri dari beberapa tahapan yaitu : Praproses akuisisi data, *input* data, seleksi data, proses penentuan program studi, split data menggunakan *k-fold cross validation*, Model *text word cloud* dan Evaluasi Model. Tahapan penelitian dapat dilihat pada gambar 1.



Gambar 1 Tahapan penelitian.

Praproses bertujuan untuk menentukan visualisasi data dengan menggunakan model *text word cloud*. Adapun tahapan praproses yaitu akuisisi data mahasiswa.

Akuisisi Data Mahasiswa, *Input* Data, Seleksi Data, Pembagian program studi, melakukan training data dan *testing data* menggunakan *k-fold cross-validation*, membuat model dengan menggunakan *text word cloud* dan evaluasi model.

Praproses

Akuisisi Data Mahasiswa

Akuisisi data mahasiswa dilakukan dengan menggunakan data sekunder dalam format excel pada bulan mei tahun 2021. Data sekunder adalah pengumpulan data melalui cara tidak langsung atau harus melakukan pencarian mendalam dahulu seperti melalui internet, literatur, statistik, buku, dan lain-lain Sugiyono (2013). Adapun *dataset* yang digunakan untuk training dan testing yaitu pertanian, pertenakan, system teknologi iformasi (STI) dan Hukum.

Input Data

Proses *Input* data merupakan langkah penting sebelum memulai proses olah data. *Input* data menggunakan *tools importing* data pada R, dengan dua cara, yaitu menggunakan *command line* dan menggunakan fasilitas *GUI R-Cmndr*. Terdapat empat perintah yang dapat digunakan yakni *read.delim, read.csv, scan* dan *data.frame* (Wickham H dan Garrett G.2017). Penelitian ini *input* data menggunakan *tools Import dataset.*

Seleksi Data

Seleksi data dilakukan untuk mengurangi variabel yang tidak relevan, variabel yang tidak relevan dapat mempengaruhi hasil visualisasi model. Penggunaan fungsi *dplyr* dan fungsi *filter* pada r studio dapat digunakan sebagai seleksi data variabel atau pemilihan responden pada *dataset* Gio dan Effendie (2017). Adapun variabel yang terdapat pada *dataset* yaitu nomor, nama, jenis kelamain, fakultas dan prodi.

Program Studi

Program studi adalah kesatuan rencana belajar sebagai pedoman penyelenggaraan pendidikan akademik dan/atau profesional yang diselenggarakan atas dasar suatu kurikulum serta ditujukan agar mahasiswa dapat menguasai pengetahuan, keterampilan, dan sikap sesuai dengan sasaran kurikulum (Kepmen PN 2000).

K-fold cross-validation

Metode *k-fold cross-validation* yaitu melakukan uji *dataset* pada program studi, dimana tahapan ini melakukan *traning data* dan *testing data* (Mahardika et al. 2017).

Teknik validasi model untuk menilai hasil statistic analisis *dataset* independen. Teknik ini digunakan untuk memecah data menjadi k bagian set data dengan ukuran yang sama, digunakan untuk melakukan prediksi model dan penggunaan *k-fold cross-validation* dapat menghilangkan bias pada *dataset* Tempola et al. 2018.

Model *Text Mining*

Metode text mining digunakan untuk mengetahui kata kunci yang sering digunakan dalam penulisan. Pengolahan data menggunakan r studio dalam analisis *text* dan visualisasi *keyword* dapat membantu kinerja peneliti. Peneliti dapat membuat model visualisasi tulisan menggunakan *text word cloud* pada r studio Zhao Y (2013).

Evaluasi Model

Pada tahapan evaluasi model menggunakan *pseudocode* pada r studio untuk mendapatkan informasi mengenai nilai matrik untuk frekuensi kata yang sering muncul pada model pada model *text word cloud* Zhao Y (2013).

III. HASIL DAN PEMBAHASAN

Akuisisi Data Mahasiswa

Akuisisi data dilakukan meng-gunakan data sekunder dalam bentuk format excel, kemudian dikonversi kebentuk format *Comma Separated Values* (CSV). Tujuan konversi guna mempermudah dalam olah *dataset*. Adapun jumlah *dataset* yaitu sebanyak 71 *testing data* dengan 5 *variabel* yaitu nomor, nama, jenis kelamin, fakultas dan program studi (Prodi). Selanjutnya *dataset* akan diolah menggunakan bahasa pemprograman r studio.

***Input* Data Ke R studio**

Data akuisisi yang telah diubah kedalam bentuk format CSV. Kemudian di *input* ke r studio dengan menggunkan *tools import dataset*. Adapun *pseudocode input dataset* di r studio ditampilkan dibawah ini.

```
library(readr)
Dataset_Mahasiswa <-
read_delim("D:/Dataset_Mahasisw
a.csv", ";" escape_double =
FALSE, trim_ws = TRUE)
View(Dataset_Mahasiswa)
```

Penggunaan *library* (readr) yaitu digunakan untuk pemanggilan atau *input* data dalam format CSV. Kemudian *Dataset*_Mahasiswa merupakan nama file data yang tersimpan pada partisi sistem D. View digunakan untuk melihat hasil tampilan dari *dataset* yang telah di *input*. Hasil dari *dataset* yang telah di *input* ke r dapat dilihat pada gambar 2.

Nomor	Nama	Jenis Kelamin	Fakultas	Prodi
1	Hartanto	L	Pertanian dan Peternakan	Pertanian
2	April	L	Hukum dan Ilmu Sosial	Hukum
3	Kurniawan	L	Teknik Ilmu Komputer	STI
4	Agustian	L	Teknik dan Ilmu Komputer	STI
5	Yulia Novita	P	Keguruan dan Ilmu Pendidikan	Keguruan
6	Arif Dermawan	L	Pertanian dan Peternakan	Pertanian
7	Fenti	P	Pertanian dan Peternakan	Pertanian
8	Kania sari	P	Hukum dan Ilmu Sosial	Hukum
9	Akhmad	L	Hukum dan Ilmu Sosial	Hukum
10	Marcheta	P	Pertanian dan Peternakan	Pertanian
11	Kemeswara	L	Keguruan dan Ilmu Pendidikan	Keguruan
12	Rangga	L	Keguruan dan Ilmu Pendidikan	Keguruan
13	Putri	P	Keguruan dan Ilmu Pendidikan	Keguruan
14	Kadir	L	Teknik dan Ilmu Komputer	STI
15	Aditya	L	Teknik dan Ilmu Komputer	STI
16	Rendy	L	Teknik dan Ilmu Komputer	STI
17	Andri	L	Teknik dan Ilmu Komputer	STI
18	Gultom	L	Pertanian dan Peternakan	Pertanian
19	Setiawan	L	Pertanian dan Peternakan	Pertanian
20	Intan Purnamasari	P	Hukum dan Ilmu Sosial	Hukum
21	Khadijah	P	Hukum dan Ilmu Sosial	Hukum

Gambar 2 Hasil *input* data ke r.

Selanjutnya untuk melihat ringkasan *dataset* atau agreget di r menggunakan *pseudocode dibawah ini*.

```
summary(Dataset_Mahasiswa)
```

Hasil dari summary dapat dilihat pada tampilan gambar 3.

```
> summary(Dataset_Mahasiswa)
     Nomor          Nama           Jenis Kelamin       Fakultas
 Min.   : 1.0   Length:71          Length:71          Length:71
 1st Qu.:18.5   Class :character   Class :character   Class :character
 Median :36.0   Mode  :character   Mode  :character   Mode  :character
 Mean   :36.0
 3rd Qu.:53.5
 Max.   :71.0
    Prodi
 Length:71
 Class :character
 Mode  :character
```

Gambar 3 Hasil *summary dataset*.

Pada gambar 4 terlihat data agreat mahasiswa, dimana nilai length sebanyak 71 dan tipe data berupa karakter.

Seleksi Data

Seleksi data pada penelitian ini bertujuan memilih variabel yang dibutuhkan dan variabel yang tidak diperlukan dihapus dengan menggunakan *pseudocode* berikut.

```
Dataset_Mahasiswa<-
data.frame(Dataset_Mahasiswa
)
Dataset_Mahasiswa<-
Dataset_Mahasiswa[-c(1,3,4)]
View(Dataset_Mahasiswa)
```

Variabel 1, 3 dan 4 pada *Dataset*_Mahasiswa berupa data frame dihapus. Sehingga hasil dari *pseudocode* tersebut dapat dilihat pada tampilan gambar 4.

Nama	Prodi
Hartanto	Pertanian
April	Hukum
Kurniawan	STI
Agustian	STI
Yulia Novita	Keguruan
Arif Dermawan	Pertanian
Fenti	Pertanian
Kania sari	Hukum
Akhmad	Hukum

Gambar 4 Hasi Seleksi data.

Hasil seleksi data terdapat 2 variabel yaitu Nama dan Prodi, variabel ini yang akan digunakan dalam visualisasi model selanjutnya.

Program Studi

Pada variabel program studi (Prodi), akan dilakukan *filter* data, yang bertujuan untuk memilih baris atau *cases* pada *data frame* dengan kondisi yang telah ditentukan. Pada *dataset* dilakukan *filter* data berdasarkan Prodi. Tampilan seperti dibawah ini.

filter(*Dataset*_Mahasiswa, Prodi=="STI")

```
> filter(Dataset_Mahasiswa, Prodi=="STI")
           Nama Prodi
1      Kurniawan   STI
2       Agustian   STI
3          Kadir   STI
4         Aditya   STI
5          Rendy   STI
```

Gambar 4 *Filter* data STI.

filter(*Dataset*_Mahasiswa, Prodi=="Pertanian")

Hasil tampilan pertanian pada gambar 5.

```
> filter(Dataset_Mahasiswa, Prodi=="Pertanian")
           Nama     Prodi
1      Hartanto Pertanian
2  Arif Dermawan Pertanian
3         Fenti Pertanian
4      Marcheta Pertanian
5        Gultom Pertanian
```

Gambar 5 *Filter* data pertanian.

Sedangkan untuk *pseudocode filter* data prodi peternakan dan hokum dapat dilihat dibawah ini dan tampilan hasil pada gambar 6 dan 7.

filter(*Dataset*_Mahasiswa, Prodi=="Peternakan")

```
> filter(Dataset_Mahasiswa, Prodi=="Peternakan")
     Nama      Prodi
1 Adiyasah Peternakan
2    tama Peternakan
3 Sungkono Peternakan
4  Gultomi Peternakan
```

Gambar 6 *Filter* data peternakan.

filter(*Dataset*_Mahasiswa, Prodi=="Hukum")

```
> filter(Dataset_Mahasiswa, Prodi=="Hukum")
              Nama Prodi
1            April Hukum
2       Kania sari Hukum
3           Akhmad Hukum
4  Intan Purnamasari Hukum
```

Gambar 7 *Filter* data hukum.

Kemudian *filter* prodi selanjutnya yaitu prodi keguruan dengan *pseudocode* sebagai berikut.

filter(*Dataset*_Mahasiswa, Prodi=="Keguruan")

```
> filter(Dataset_Mahasiswa, Prodi=="Keguruan")
          Nama    Prodi
1  Yulia Novita Keguruan
2     Kemeswara Keguruan
3        Rangga Keguruan
4         Putri Keguruan
5         Peter Keguruan
```

Gambar 8 *Filter* data keguruan

Hasil dari *filter* data keseluruan prodi didapatlah 5 prodi yaitu STI, Pertanian, Peternakan, Hukum dan Keguruan dengan jumlah total *dataset* 71 *testing data* .

K-fold cross-alidation

Tahapan *k-fold cross-validation* pada penelitian ini dengan *training data* dan *testing data* k=10, maka didapatlah *training* 54 *testing data* dengan jumlah variabel 2 dan *testing data* sebanyak 17 *testing data* dengan 2 variabel.

Selanjutnya membuat viualisasi model text mining denan mengunakan *training data* dan *testing data*.

Model *Text Mining*

Model *text mining* pada penelitian ini menggunakan visualisasi data *word cloud*, *pseudocode* sebagai berikut.

```
wordcloud(words = d$word, freq
= d$freq, min.freq =
1max.words=200,
random.order=FALSE,
rot.per=0.35,
            colors=brewer.pal(8,
"Dark2"))
```

Hasil dari tampilan *pseudocode* menggunakan *training data* diatas data dilihat pada gambar 9.

Gambar 9 *Word cloud training data.*

Gambar 9 *Word cloud training data* diatas menggunakan 54 *testing data* dan menunjukan bahwa kata "STI", "Keguruan", "Pertanian", "Hukum" dan "Peternakan" adalah lima kata terpenting yang sering digunakan dalam *dataset* mahasiswa. Selanjutnya pada gambar 10 merupakan tampilan *word cloud* pada *testing data.*

Gambar 10 Tampilan *word cloud* pada *testing data.*

Tampilan *word cloud* pada gambar 10 dengan menggunakan *testing data* menunjukan bahwa terdapat 4 kata terpenting, diantaranya "STI", "Keguruan",

"Hukum" dan "Pertanian". Penggunaan data testing sebanyak 17 *testing data*. Hasil model *testing data* dalam betuk visualisasi *word cloud* digunakan sebagai analisis dan evaluasi model

Evaluasi Model

Evaluasi terhadap *testing data* divisualisasikan ke tampilan *word cloud* yaitu dengan jumlah data 17 *testing data* dengan prodi "sti" nilai frekuensi 5, "keguruan" nilai frekuesi 4, "hukum" nilai frekuensi 3 dan nilai frekuensi "pertanian" 2. Sehingga pengolahan data sederhana dengan menggunakan Bahasa pemprogram r studio dapat digunakan dan divisualisaikan dengan jelas.

IV. SIMPULAN

Pengolahan data sederhana menggunakan r studio dalam dapat di visulisasikan dengan baik menggunakan tampilan *Word cloud*. *Dataset* yang digunakan pada penelitian ini sebanyak 71 *record* dengan jumlah variabel 5, kemudian dilakukan seleksi data, sehingga didapatlah jumlah variabel 2 dan *dataset* 71 *record*. Data tersebut divisualisasikan ke betuk word cloud, kemudian penggunan *testing data* sebagai evaluasi data. Pengolahan data serdehana dengan menggunakan r studio guna mempermudah peneliti dalam olah data.

DAFTAR RUJUKAN

Budiharto W dan Ro'fah Nur Rachmawati. (2013). *Pengantar Praktis Pemrograman R untuk Ilmu Komputer*. Jakarta: Moeka.

Gio P dan Adhitya RE. (2017). *Belajar Bahasa Pemrograman R(Dilengkapi Cara Membuat Aplikasi Olah Data Sederhana dengan R Shiny)*. Medan: USU Press.
Wickham H dan Garrett G. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data 1st Edition*. Newton : O'Reilly.

Keputusan Menteri Pendidikan Nasional Republik Indonesia Nomor 232/U/2000 Tentang Pedoman Penyusunan Kurikulum Pendidikan Tinggi Dan Penilaian Hasil Belajar Mahasiswa.

Mahardika YM, Sudarsono A, Barakbah AR. (2017). An Implementation Of Bonet Dataset to Predicate Accuracy based On Network Flow Model. International Electric Symposiumon Knowladge Creation and Intelegance Computing.

Peraturan Dikti Keputusan Menteri Pendidikan Nasional Nomor 232/U/2000 12 Februari 2006, 23:34:08.

Permana S, Yuniarto B, Mariyah S, Ibnu S, dan Nooraeni R. (2018). Data Mining Dengan R Konsep Serta Implementasi. Bogor: In Media

Setiawan N. (2005). *Pengolahan dan Analisis Data. Bandung* : Universitas Padjadjaran.

Sugiyono. (2013). *Metode Penelitian Bisnis.* Bandung: Alfabeta.

Sugiyono.(2010). *Metode Penelitian Kuantitatif, kualitatif dan R & D*. Bandung: Alfabeta.

Tempola F, Miftah M, dan Amal K. (2018). Perbandingan Klasifikasi Antara Knn dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K. *Jurnal Teknologi Informasi dan Ilmu Komputer* (JTIIK). 5.5: 577-584.

Zhao, Y. 2012. *R and Data Mining: Examples and Case Studies*. Leiden: Elsevier .