

# Analisis Sistem Informasi Pendistribusian Tempat KKL Menggunakan Metode *K-Nearst Neighbor*

Hasrati Mayangsari<sup>1</sup>, Dewi Kartika<sup>2</sup>, Dr.Eng.Wiliem Musu,S.Kom.,M.T<sup>3</sup>, Heriadi,S.Pd.I,M.Pd<sup>4</sup>

<sup>1,2</sup> Jurusan Sistem Informasi Universitas Dipa Makassar  
Jln. Perintis Kemerdekaan KM. 9 Makassar

<sup>1</sup>hasratimayangsari06@gmail.com, <sup>2</sup>dewikartika@gmail.com, <sup>3</sup>wilem.musu@undipa.ac.id, <sup>4</sup>Heriadi@undipa.ac.id

## Abstrak

Penelitian ini berfokus pada pemncarian model prediksi dalam pendistribusian mahasiswa KKL ke intansi KKL dengan menggunkana metode prediski *K-Nearst Neighbor*. Dengan menguji performa dari dataset hasil kuesioner dengan jumlah 225 data, dengan hasil *accuracy* yang diperoleh hanya 41%. Hasil yang diperoleh menunjukkan model tidak dapat digunakan untuk masa yang mendatang, sehingga dataset hasil kuesioner harus di *clustering* ulang berdasarkan karakteristik yang sama dengan menggunakan algoritma *K-Means*. Hasil dari algoritma *K-measn* akan di uji kembali menggunakan metode *KNN*, dengan hasil nilai akurasi yang diperoleh yaitu 91%, yang artinya bahwa model yang bagus dalam memprediksi pendistribusian mahasiswa KKL adalah model yang dihasilkan dari *clustering K-Means* kemudian di prediksi menggunakan metode *KNN*.

**Kata Kunci :** *K-Nearst Neighbor* dan *K-Means*.

## I. PENDAHULUAN

KKL merupakan kepanjangan dari Kuliah Kerja Lapangan yang dimana merupakan proses perkuliahan dalam bentuk praktek kerja dilapangan dalam sebuah perusahaan atau kantor untuk menerapkan pengetahuan dan teori yang telah didapatkan selama proses perkuliahan di kampus. Untuk itu mahasiswa diwajibkan mengikuti program KKL dengan syarat yang telah ditentukan oleh pihak kampus.

Dalam melaksanakan program KKL, mahasiswa diharuskan untuk mencari perusahaan atau kantor yang bisa dijadikan tempat KKL. Dalam proses pencarian tempat KKL, mahasiswa lebih mengandalkan orang dalam seperti keluarga yang bekerja didalam perusahaan ketimbang mencari terlebih dahulu tempat KKL yang cocok dengan kemampuannya. Dan bagi mahasiswa yang tidak tau lokasi tempat KKL, akan cenderung mengikuti mahasiswa yang memiliki pengaruh berupa orang dalam yang bekerja didalam perusahaan, tanpa berpikir panjang. Sehingga banyak mahasiswa yang mengalami ketidakmampuan atau ketidak cocokan dalam bekerja di dalam perusahaan yang tidak sesuai dengan kemampuan yang dimilikinya. Berdasarkan hal tersebut, proses pelaksanaan KKL mahasiswa akan tidak berjalan sesuai dengan tujuan dari dilakukannya program KKL, karena ketidaktahuan mahasiswa dalam memilih tempat KKL yang sesuai dengan kemampuan yang dimilikinya.

Data mitra tempat KKL merupakan sebuah sumber informasi yang dapat dijadikan rujukan untuk memilih tempat KKL, yang dimana memuat informasi dari data mitra tempat KKL yang meliputi instansi pemerintahan, BUMN, BUMD, dan perusahaan swasta dimana digunakan sebagai acuan dalam penempatan mahasiswa KKL.

Untuk menempatkan mahasiswa keinstansi yang sama dengan kemampuan yang dimiliki mahasiswa dan sesuai kebutuhan instansi perusahaan, maka dibutuhkan kajian

kesesuaian kebutuhan tempat KKL dan kemampuan mahasiswa yang melakukan KKL. Untuk itu diperlukan data dari mahasiswa, dengan cara pengumpulan data berupa kuesioner yang akan dibagikan ke 225 mahasiswa KKL. Setelah itu data akan dianalisis menggunakan *Regresi Linear* untuk melihat tingkat kedekatan antara variabel, dengan tujuan untuk mengetes model dari *K-NN*, sedangkan *K-Means* berguna untuk menyesuaikan dan melengkapi hasil dari seleksi klasifikasi dari metode *K-NN*.

Tujuan dari penelitian ini adalah untuk menentukan model prediksi yang sesuai dalam pendistribusian mahasiswa ke instansi KKL, dengan menggunakan metode *K-Nearst Neighbor (KNN)* sebagai metode prediksi, *Multiple Regresion Linear* sebagai analisis antar korelasi variabel dan metode *K-Means* yang digunakan untuk mengelompokkan data dengan karakteristik.

### A. Data Mining

Secara sederhananya data mining mengarah pada ekstraksi dan pertambahan pengetahuan dari sejumlah besar data. [1]

Data mining merupakan suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa didalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika. [2]

Data mining terbagi menjadi beberapa kelompok berdasarkan tugas yang dapat di lakukan, antara lain [3]:

#### 1) Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, peng-golongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.

2) Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model di bangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, berat badan, dan level sodium darah.

3) Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Contoh prediksi dalam bisnis yaitu, Prediksi harga beras dalam tiga bulan yang akan datang.

B. K-Neasrt Neighbor (KNN)

K-Nearest Neighbor (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran (*neighbor*) yang jaraknya paling dekat dengan objek tersebut. Dekat atau jauhnya *neighbor* biasanya dihitung berdasarkan jarak *Euclidean*. diperlukan suatu sistem klasifikasi sebagai sebuah sistem yang mampu mencari informasi [4]. Metode KNN dibagi menjadi dua fase, yaitu pembelajaran (training) dan klasifikasi atau pengujian (testing). Pada fase pembelajaran, algoritma ini hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi dari data pembelajaran. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk data yang akan diuji coba (yang klasifikasinya tidak diketahui). Jarak dari vektor yang baru ini terhadap seluruh vektor data pembelajaran dihitung, dan sejumlah k buah neighbor yang paling dekat diambil. Perhitungan jarak ketetanggan menggunakan algoritma eucliden seperti yang ditunjukkan pada persamaan 1.

$$d_{Euclidean}(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \tag{1}$$

C. Multiple Regression Berganda

Analisis regresi linier adalah metode *statistic* yang dapat digunakan untuk mempelajari hubungan antar sifat permasalahan yang sedang diselidiki. Model regresi linier berganda merupakan model regresi yang melibatkan lebih dari satu variable bebas. Analisis regresi linier berganda bertujuan untuk mengetahui arah dan seberapa besar pengaruh variable bebas terhadap variable terikat.

$$y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n \tag{2}$$

Pada persamaan 1 variabel y merupakan variable terikat, a merupakan intersept atau konstanta regresi, x merupakan variable bebas, dan b merupakan koefisien regresi pada masing-masing variable bebas. Besarnya keterkaitan antara variable terikat dengan variable bebas dinyatakan oleh koefisien korelasi.

$$r = \frac{N \sum_{i=1}^N (x_i y_i) - \sum_{i=1}^N (x_i) \cdot \sum_{i=1}^N (y_i)}{\sqrt{[N \sum_{i=1}^N (x_i^2) - (\sum_{i=1}^N (x_i))^2][N \sum_{i=1}^N (y_i^2) - (\sum_{i=1}^N (y_i))^2]}} \tag{3}$$

Nilai r yang positif mengindikasikan bahwa meningkatnya nilai x akan menyebabkan meningkatnya nilai y, sebaliknya untuk nilai r yang negative mengindikasikan bahwa meningkatnya nilai x akan menyebabkan menurunnya nilai y, dan jika nilai r sama dengan nol mengindikasikan bahwa tidak ada korelasi antara x dan y. Pengujian model algoritma yang digunakan dapat dilakukan dengan cara mencari nilai *Mean Square Error (MSE)*, *Root Mean Square Error (RMSE)*, dan *Mean Absolute Error (MAE)*. *MSE* merupakan kuadrat rata-rata dari nilai error [5].

$$MSE = \sum \frac{(y' - y)^2}{n} \tag{4}$$

dimana y' merupakan nilai variable terikat hasil prediksi, y merupakan nilai variable terikat sebenarnya, dan n merupakan jumlah data. *Root Mean Square Error (RMSE)* adalah jumlah dari kesalahan kuadrat atau selisih antara nilai sebenarnya dengan nilai prediksi yang telah ditentukan.

$$RMSE = \sqrt{\sum \frac{(y' - y)^2}{n}} \tag{5}$$

*Mean Absolute Error (MAE)* menunjukkan nilai kesalahan rata-rata yang error dari nilai sebenarnya dengan nilai prediksi. *MAE* sendiri secara umum digunakan untuk pengukuran prediksi error pada analisis time series.

$$MAE = \sum \frac{|y' - y|}{n} \tag{6}$$

D. K-Means

*K-Means* digunakan untuk meminimalisasikan fungsi objective yang telah di set dalam proses *clustering* sebelumnya. Tujuan tersebut dilakukan dengan cara meminimalikan variasi data yang ada didalam *cluster* dan memaksimalikan variasi data yang ada di *cluster* lainnya.

*K-Means Clustering* adalah suatu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode *K-Means* berusaha mengelompokkan data yang ada kedalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada didalam kelompok yang lain. Dasar algoritma *K-Means* adalah sebagai berikut : [6]

1. Tentukan k sebagai jumlah *cluster* yang akan dibentuk
2. Tentukan k *Centroid* (titik pusat *cluster*) awal secara acak. Adapun rumus yang digunakan untuk menghitung *centroid cluster* ke-i berikutnya sebagai berikut:

$$k = \frac{\sum_{i=1}^n X_i}{n}; i = 1, 2, 3, \dots, n \quad (7)$$

Dimana :

k = centroid pada *cluster*

$X_i$  = objek ke-i

n = banyaknya jumlah

3. Hitung jarak setiap objek ke masing-masing *centroid* dari masing-masing *cluster*. Untuk menghitung jarak antara objek dengan *centroid* dapat menggunakan Euclidean Distance.

$$d(x,y) = \| x - y \| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}; I = 1'2'3' \dots, n \quad (8)$$

Dimana :

$x_i$  = objek x ke-i

$y_i$  = data y ke-i

n = banyaknya objek

4. Alokasikan masing-masing objek ke dalam *centroid* yang paling dekat.
5. Lakukan iterasi, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan (8)
6. Ulangi langkah 3 jika posisi *centroid* baru tidak sama.[7]

#### E. Confusion matriks

Confusion matrix merupakan sebuah model evaluasi klasifikasi berdasarkan data uji dan seluruh data yang diprediksi dengan proporsi yang tepat [8]. Nilai *confusion matriks* terdiri dari nilai *accuracy*, *precision*, *recall*, dan *f1 measure*. Dimana *Accuracy* didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual. rumus *accuracy* dipaparkan pada persamaan 5. *Precision* didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. *Precision* dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. rumus *precision* ditunjukkan pada persamaan 10. *Recall* didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. Rumus *Recall* diuraikan pada persamaan 11, dan F-Measure adalah harmonic mean antara nilai *precision* dan *recall*, F-measure juga kadang disebut dengan nama *F1-score*. Rumus F-Measure dijabarkan pada persamaan 12.

$$ACCURACY = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (9)$$

$$PRECISION = \frac{TP}{TP + FP} \quad (10)$$

$$RECALL = \frac{TP}{TP + FN} \quad (11)$$

$$F - \text{Measure} = 2 \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \quad (12)$$

#### F. Python

Python merupakan salah satu bahasa pemrograman yang banyak digunakan oleh perusahaan besar maupun para developer untuk mengembangkan berbagai macam aplikasi berbasis desktop, web dan mobile. Python diciptakan oleh Guido van Rossum di Belanda pada tahun 1990 dan namanya diambil dari acara televisi kesukaan Guido Monty Python's *Flying Circus*. Van Rossum mengembangkan Python sebagai hobi, kemudian Python menjadi bahasa pemrograman yang dipakai secara luas dalam industri dan pendidikan karena sederhana, ringkas, sintak *sintuitif* dan memiliki pustaka yang luas[9].

#### G. Penelitian Terkait

Penelitian-penelitian yang dilakukan umumnya berfokus pada bagaimana mendapatkan mahasiswa ke instansi KKL yang sesuai dengan kemampuan mahasiswa. Salah satunya adalah penelitian [10] yang berjudul "*Grouping Students Academic Performance using One-Way Clustering*". Penelitian yang dilakukan pada sekolah dasar di Selangor Malaysia ini bertujuan untuk membandingkan kinerja dari metode *hierarchical clustering* dan *K-Means clustering* dalam pembentukan kelas siswa serta melihat karakteristik siswa dari kelas yang berhasil dibentuk. Dari hasil penelitian yang dilakukan diperoleh kesimpulan bahwa metode *K-Means* lebih tepat digunakan untuk mengelompokkan siswa dan menentukan karakter kelompok siswa tersebut.

Penelitian serupa sebelumnya telah dilakukan oleh [11] yang berjudul "Pembagian Kelas Kuliah Mahasiswa Menggunakan Pengklasteran *Fuzzy*". Penelitian yang dilakukan menggunakan studi kasus Jurusan Teknik Elektro Universitas Diponegoro. Penelitian yang dilakukan menunjukkan bahwa algoritma *fuzzy k-means* yang digunakan dapat membagi kelas kuliah mahasiswa lebih baik dibandingkan dengan algoritma subtraktif. Algoritma subtraktif menghasilkan *cluster* yang sulit diukur tingkatan kemampuan siswa di dalamnya.

Penelitian lain juga dilakukan [12] yang berjudul "*Clustering Student for Group Based Learning in Foreign Language Learning*". Penelitian yang dilakukan berfokus pada mekanisme pembentukan kelompok siswa otomatis dengan memanfaatkan metode *clustering* untuk pembelajaran bahasa asing. Pada penelitian ini digunakan dua jenis strategi pengelompokan siswa, yaitu *similarity clustering* dan *complementation clustering*. Hasil penelitian yang dilakukan menunjukkan bahwa metode yang dihasilkan lebih fleksibel dan komprehensif dibandingkan dengan pengelompokan secara manual dalam pembelajaran bahasa asing

## II. METODOLOGI PENELITIAN

Penelitian ini dilakukan di kampus Universitas Dipa Makassar dibagian Perencanaan, Pengembangan dan Kerjasama (PPK), yang beralamat Jln Perintis Kemerdekaan KM 9 Makassar.

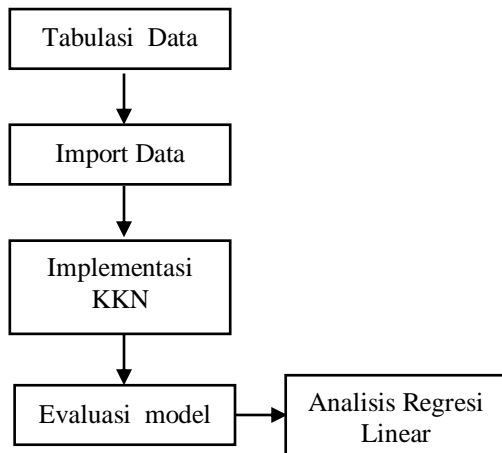
Pada proses pengujian metode, peneliti menggunakan bahasa pemrograman *python*. Dengan melakukan pengujian

metode *K-Nearst Neighbor (KNN)* dalam membangun model prediksi, pengujian anova digunakan untuk melihat hubungan atau pengaruh antar variabel, sedangkan metode *K-Means*, digunakan untuk mengelompokkan data berdasarkan karakteristik yang sama.

Adapun tahapan proses pengujian model *KNN* dengan *pyhton* dapat dilihat pada gambar berikut.

1. Tahap pertama pengujian *KNN*

Pada tahap ini dataset yang akan diuji merupakan dataset hasil kuesioner dengan jumlah 255, dengan pengujian metode *KNN* sebagai *pyhton* adalah sebagai berikut.

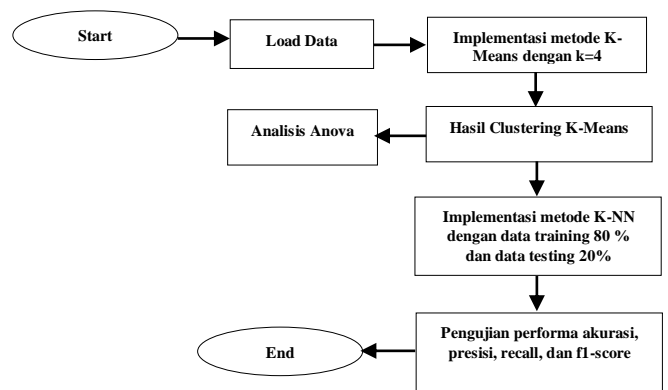


Gambar 1. Tahapan Proses Pengujian *KNN*

- a. Tabulasi Data  
Data hasil kuesioner akan ditabulasi menjadi data yang bisa dianalisis.
- b. Import Data  
Data yang bisa dibaca atau dianalisis di *pyhton* adalah data dengan tipe *csv*.
- c. Implementasi *KNN*  
*KNN* merupakan metode yang bisa melakukan klasifikasi terhadap objek berdasarkan data training dan data testing. dimana data training sebanyak 80% dan data testing sebanyak 20%.
- d. Evaluasi Model *KNN*  
*Confusion matriks* akan digunakan dalam penentuan performa pada model prediksi *KNN*.
- e. Analisis Regresi Linear Berganda  
Regresi linear berganda merupakan model regresi yang melibatkan lebih dari satu variabel independen. Analisis regresi linear berganda dilakukan untuk mengetahui arah dan seberapa besar pengaruh variabel independen terhadap variabel dependen [13].

2. Tahap pengujian Metode *KNN* dengan metode *K-Means*

Tahap ini dilakukan untuk mendapatkan model prediksi *KNN* yang cocok dalam pendistribusian mahasiswa ke instansi *KKL*. Dengan tahapan sebagai berikut.



Gambar 2. Alur tahapan pengujian *KNN* dan *K-Means*

- a. Import Data  
Dalam *pyhton*, Pengimputan file dataset dalam *pyhton* harus ber-format *csv*. *csv* merupakan *Comma Separated Values* atau *CSV* adalah suatu format data dalam basis data di mana setiap *record* dipisahkan dengan tanda koma (,) atau titik koma.
- b. Implementasi metode *K-Means*  
*K-Means Clustering* adalah metode untuk mengkategorikan atau pengelompokan sekelompok objek sesuai dengan atribut yang sama atau karakteristik ke dalam sejumlah groups. Dimana clustering(grub) yang dihasilkan sebanyak 4, yang terdiri dari instansi pemerintahan, *BUMN*, *BUMD*, dan perusahaan swasta.
- c. Hasil clustering *K-Means*  
Perolehan dari *k-means* akan di analisis menggunakan *anova* atau *Analysis of variance*. dengan analisis satu arah, yang dimana merupakan pengujian dengan memakai varian serta data hasil pengamatan terhadap satu faktor.
- d. Implementasi Metode *KNN*  
*K-Nearst Neighbor* atau *KNN* merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.  
*KNN* memiliki keunggulan untuk mengklasifikasikan data mahasiswa yang tidak diketahui dengan adanya data latih dan data uji. *KNN* dapat menprosedur yang berbasis matematis untuk mengevaluasi nilai kriteria-kriteria tersebut menjadi sebuah keterangan klasifikasi.
- e. Performa model prediksi *KNN*  
Untuk melihat performa model *KNN* yang diuji sebelumnya, maka dilakukan perhitungan *confusin matrix* yang terdiri dari nilai akurasi, *precision*, *recall*, dan *f1-score*. Dimana jika nilai akurasi yang dihasilkan sangat tinggi, maka model dikatakan dapat digunakan di masa yang mendatang, sedangkan jika nilai yang

dihasilkan rendah, maka model yang dihasilkan sangat buruk dan tidak bisa digunakan untuk prediksi masa depan.

### III. HASIL DAN PEMBAHASAN

Adapun langkah langkah pengujian metode dengan menggunakan bahasa pemrograman *python* adalah sebagai berikut :

#### 1. Pengujian Model KNN

Pada tahap pengujian pertama merupakan pengujian metode KNN dalam menentukan performa model prediksi KNN yang terbaik dalam pendistribusian mahasiswa ke instnasi KKL menggunakan data hasil tabulasi kuesioner. Dengan langkah-langkah sebagai berikut:

a) Import liblary yang dibutuhkan

Pengimputan liblary pada *pyhton* dapat membuat Pemrograman *Python* lebih sederhana dalam proses pengkodingan.

```
import numpy as np
import pandas as pd |
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
```

Gambar 1. Import liblary KNN

b) import dataset

Pemanggilan data dalam file dengan format csv.

```
#pemanggilan dataset dari file excel
dataset = pd.read_csv('dataset.csv', sep=';')
dataset
```

Gambar 2. Import Dataset

c) Split Data

Split data merupakan pembagian data training dan data testing.

```
X = dataset.iloc[:,0:4]
y = dataset.iloc[:,4]
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, test_size=0.2)
```

Gambar 3. Split Data

d) Normalisasi Data

Normalisasi data dilakukan dapam proses algoritma adalah untuk membuat beberapa variabel memiliki rentang nilai yang sama, tidak ada yang terlalu besar maupun terlalu kecil sehingga dapat membuat analisis statistik menjadi lebih mudah.

```
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
```

Gambar 4. Normalisasi Data

e) Implementasi metode KNN

```
knn_model = KNeighborsClassifier()
knn_model.fit(X_train, y_train)
KNeighborsClassifier()
```

Gambar 5. Implementasi metode KNN

f) Evaluasi Model KNN

```
: print(classification_report(y_test, knn_model.predict(X_test)))
```

|                     | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| BUMN                | 0.22      | 0.20   | 0.21     | 10      |
| Instansi Pemerintah | 0.61      | 0.67   | 0.63     | 30      |
| Perusahaan Swasta   | 0.00      | 0.00   | 0.00     | 5       |
| accuracy            |           |        | 0.49     | 45      |
| macro avg           | 0.28      | 0.29   | 0.28     | 45      |
| weighted avg        | 0.45      | 0.49   | 0.47     | 45      |

Gambar 6. Evaluasi model KNN

Hasil akurasi yang ditunjukkan pada pengujian KNN menggunakan dataset hasil kuesioner menunjukkan nilai 41% , yang berarti bahwa model prediksi dari hasil kuesioner tidak layak digunakan untuk masa medatang.

#### 2. Analisis Regresi

Analisis regresi linear berganda dilakukan untuk mengetahui arah dan seberapa besar pengaruh variabel independen terhadap variabel dependen, dengan langkah - langkah sebagai berikut :

a) Import *liblary* pada *python*

Pengimputan liblary yang dilakukan pada *pyhton* dapat membuat Pemrograman *Python* lebih sederhana.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib as mpl
import scipy.stats as stats
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score
import math
```

Gambar 7. Liblary Regresi

b) Import dataset

```
dataset = pd.read_csv('dataset.csv', sep=';')
dataset
```

Gambar 8. Import Dataset

c) Mengubah tipe data

Dalam analisis regresi dalam proses ananlisis semua data pada kolom harus bertipe numerik.

```
enc = LabelEncoder()
dataset['Instansi'] = enc.fit_transform(dataset['Instansi'].values)

dataset
```

Gambar 9. Mengubah Type Data

d) Split Data

Split data merupakan pembagian data training dan data testing.

```
X = dataset.iloc[:,0:4]
y = dataset.iloc[:,4]

#SPLIT DATA ANTARA TRAINING DAN TESTING DENGAN RASIO 80:20
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, test_size=0.2)
```

Gambar 10. Split Data

e) Analisis Regresi

Intercept : 2.1202363049012938

|   | Instansi     | coeficients |
|---|--------------|-------------|
| 0 | Administrasi | -0.022038   |
| 1 | Komunikasi   | 0.012437    |
| 2 | Pemrograman  | 0.014100    |
| 3 | Multimedia   | 0.005995    |

Gambar 11. Analisis Regresi

Berdasarkan hasil coeficients, maka dapat disimpulkan bahwa variabel x (administrasi, komunikasi, pemrograman, dan multimedia) tidak memiliki kontribusi terhadap keempat variabel.

f) Evaluasi Model Regresi

```
y_predic = gresi.predict(X_test)

print('Mean Absolute Error (MAE): %.2f' % mean_absolute_error(y_predic, y_test))
print('Mean Squared Error (MSE): %.2f' % mean_squared_error(y_predic, y_test))
print('Root Mean Squared Error (RMSE): %.2f' % math.sqrt(mean_squared_error(y_predic, y_test)))
print('R2-score: %.2f' % r2_score(y_predic, y_test))

Mean Absolute Error (MAE): 0.38
Mean Squared Error (MSE): 0.33
Root Mean Squared Error (RMSE): 0.57
R2-score: -35.86
```

Gambar 12. Evaluasi model Regresi

Perhatikan nilai R2-score yaitu -35.86. Hasil tersebut menunjukkan bahwa tidak ada pengaruh antar variabel independen terhadap variabel dependen.

Dengan ini, nilai evaluasi regresi dapat membuktikan bahwa hasil prediksi KNN menggunakan dataset dari hasil kuesioner sangat tidak layak digunakan.

3. Pengujian metode K-Means dengan Metode KNN

Pengujian ini dilakukan untuk meningkatkan performa model KNN dengan melakukan Clustering dari hasil K-Means, adapun langkah – langkah pengujian pada pyhton adalah sebagai berikut :

a) Import *library* yang dibutuhkan

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from scipy.spatial.distance import cdist
import matplotlib.pyplot as plt
```

Gambar 13. Import *library* K-Means

b) Import dataset

```
data = pd.read_csv('dataset.csv', sep=';')
dataset = data.iloc[:,0:4]
dataset
```

Gambar 14. Pemanggilan dataset

c) Normalisasi data

```
Scaler = MinMaxScaler()
X_train = Scaler.fit_transform(X_train)
X_train
```

Gambar 15. Normalisasi Data

d) Membuat model *cluster*

```
klaster = KMeans(n_clusters=4)
klaster

KMeans(n_clusters=4)
```

Gambar 16. Menentukan jumlah *cluster*

e) Evaluasi hasil *K-Means clustering*

```
#Label instansi yang diprediksi
hasil = klaster.fit_predict(X_train)
hasil

array([1, 3, 3, 3, 0, 1, 1, 1, 0, 2, 1, 1, 0, 0, 3, 1, 0, 0, 0, 3, 1, 3,
       3, 2, 3, 3, 3, 1, 1, 1, 2, 0, 1, 1, 1, 1, 0, 3, 0, 0, 1, 0, 1, 1,
       1, 3, 0, 1, 1, 1, 2, 2, 0, 1, 1, 2, 2, 3, 1, 1, 1, 0, 0, 1, 3,
       0, 1, 1, 1, 1, 0, 2, 1, 1, 0, 0, 3, 1, 1, 3, 1, 3, 0, 1, 1, 1,
       1, 0, 3, 3, 3, 2, 3, 1, 0, 1, 1, 0, 1, 1, 0, 3, 0, 1, 3, 2, 2, 3,
       0, 2, 1, 1, 0, 3, 3, 0, 2, 1, 1, 3, 0, 0, 1, 1, 3, 1, 0, 1, 1,
       1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 3, 2, 3, 3, 1, 3, 3, 1, 3, 1, 0, 1,
       2, 3, 0, 1, 1, 1, 1, 1, 1, 3, 2, 3, 2, 1, 0, 1, 3, 0, 1, 0, 0,
       0, 2, 3, 0, 3, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 2, 1, 0, 3, 3,
       0, 0, 0, 1, 3, 0, 1, 3, 1, 3, 1, 1, 0, 1, 1, 1, 2, 2, 3, 3, 1,
       0, 0, 3, 1, 1])
```

Gambar 17. Hasil *clustering* K-Means

f) Analisis Anova

Analisis anova digunakan untuk mengetahui seberapa besar pengaruh variabel dependen terhadap variabel independen, dengan hasil sebagai berikut.

| Admistrasi  |              |       |           |          |
|-------------|--------------|-------|-----------|----------|
|             | sum_sq       | df    | F         | PR(>F)   |
| Instansi    | 929.109706   | 1.0   | 8.483186  | 0.003949 |
| Residual    | 24423.779183 | 223.0 | NaN       | NaN      |
| Komunikasi  |              |       |           |          |
|             | sum_sq       | df    | F         | PR(>F)   |
| Instansi    | 171.205082   | 1.0   | 4.065729  | 0.044961 |
| Residual    | 9390.377140  | 223.0 | NaN       | NaN      |
| Pemrograman |              |       |           |          |
|             | sum_sq       | df    | F         | PR(>F)   |
| Instansi    | 102.315933   | 1.0   | 1.352352  | 0.24611  |
| Residual    | 16871.684067 | 223.0 | NaN       | NaN      |
| Multimedia  |              |       |           |          |
|             | sum_sq       | df    | F         | PR(>F)   |
| Instansi    | 1446.034804  | 1.0   | 18.743417 | 0.000023 |
| Residual    | 17204.214085 | 223.0 | NaN       | NaN      |

Gambar 18. Analisis Anova

Perhatikan nilai PR(>F) yang menjadi patokan pengaruh antar variabel dimana pengaruh variabel administrasi terhadap variabel instansi memiliki pengaruh yang sangat signifikan termasuk variabel komunikasi dan multimedia. Sedangkan variabel pemrograman memiliki nilai 0.246 yang dimana lebih besar dari tingkat signifikansi yang digunakan untuk menghitung tingkat kepercayaan yaitu 0.05. Sehingga kemampuan pemrograman tidak memiliki pengaruh terhadap variabel instansi.

g) Implementasi *K-Nearest Neighbor*

Adapun langkah-langkah pengujian metode *KNN* menggunakan hasil *clustering K-Means* pada *python* adalah sebagai berikut:

- 1) Import *library* yang dibutuhkan

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
```

Gambar 19. Import Library *KNN*

- 2) Melakukan pembagian data testing dan data training

```
X = dataset.iloc[:,0:4]
y = dataset.iloc[:,4]
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, test_size=0.2)
```

Gambar 20. Split Data

- 3) implementasi metode *KNN*

```
knn_model = KNeighborsClassifier()
knn_model.fit(X_train, y_train)

KNeighborsClassifier()

#predict the test set results

y_pred = knn_model.predict(X_test)
y_pred
array([[2, 1, 3, 0, 0, 3, 3, 1, 3, 3, 2, 3, 2, 0, 3, 0, 2, 1, 1, 1, 0, 1,
       0, 0, 3, 1, 1, 3, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 3, 3, 1, 3,
       0]])
```

Gambar 21. Implementasi metode *KNN*

- 4) Evaluasi model

tahap ini merupakan tahap *confusion matriks* yang terdiri dari nilai *accuracy*, *precision*, *recall*, dan *f1-score*, dengan tabel *confusion matriks* sebagai berikut.

```
cm = confusion_matrix(y_test, y_pred)
print(cm)

[[12  0  0  0]
 [ 0 15  0  2]
 [ 0  0  4  0]
 [ 0  2  0 10]]
```

Gambar 22. tabel *confusion matriks*

Dengan hasil evaluasi model sebagai berikut :

```
print(classification_report(y_test, knn_model.predict(X_test)))

              precision    recall  f1-score   support

0               1.00         1.00         1.00         12
1               0.88         0.88         0.88         17
2               1.00         1.00         1.00          4
3               0.83         0.83         0.83         12

 accuracy - macro avg 0.93   0.93   0.91         45
 accuracy - weighted avg 0.91   0.91   0.91         45
```

Gambar 23. Evaluasi model

*Accuracy* yang dihasilkan dari hasil *clustering K-Means* adalah 91% dimana nilai instansi pemerintah menunjukkan nilai *precision*, *recall*, dan *f1-score* = 1.00. Nilai BUMD untuk *precision*, *recall* dan *f1-score* = 0.83, perusahaan swasta menunjukkan nilai *precision*, *recall*, dan *f1-score* = 0.88 dan instansi BUMN untuk nilai *precision*, *recall*, dan *f1-score* = 1.00.

IV. KESIMPULAN

Berdasarkan hasil pengujian metode menggunakan *python* dapat disimpulkan hal-hal sebagai berikut.

- 1) Dengan menggunakan algoritma *K-Means* dapat menghasilkan klasifikasi karakteristik berdasarkan kemampuan dari mahasiswa.
- 2) Model prediksi yang sesuai dalam pendistribusian mahasiswa ke instansi KKL adalah model yang dihasilkan dari klasifikasi *K-Means* dan di uji kembali menggunakan metode prediksi *KNN*.

V. SARAN

Untuk penelitian berikutnya dapat dilanjutkan dengan menambahkan pertanyaan-pertanyaan yang berhubungan dengan bidang pekerjaan pada perusahaan atau kantor. dan dapat melakukan penambahan variabel independen (variabel kemampuan). Dan dapat diimplementasikan dalam bentuk aplikasi yang dapat dipergunakan oleh kampus Universitas Dipanegara.

REFERENSI

[1] Han, J., & Kamber, M. (2006). Classification and prediction. *Data mining: Concepts and techniques*, 2006, 347-50.

- [2] Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining* (Vol. 4). John Wiley & Sons.
- [3] Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining* (Vol. 4). John Wiley & Sons.
- [4] Lestari, M. E. I. (2015). *Penerapan algoritma klasifikasi Nearest Neighbor (K-NN) untuk mendeteksi penyakit jantung*.
- [5] Kusuma, W. S., & Sutapa, P. (2020). Dampak pembelajaran daring terhadap perilaku sosial emosional anak. *Jurnal Obsesi: Jurnal Pendidikan Anak Usia Dini*, 5(2), 1635-1643.
- [6] Ong, J. O. (2013). Implementasi algoritma *K-Means clustering* untuk menentukan strategi marketing president university.
- [7] Ediyanto, M. N. M., & Satyahadewi, N. (2013). Pengklasifikasian Karakteristik Dengan Metode *K-Means Cluster Analysis*. *Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya*, 2(02).
- [8] Faiza, N. N. (2009). Prediksi tingkat keberhasilan mahasiswa tingkat I IPB dengan metode k-Nearest Neighbor.
- [9] M. Romzi and B. Kurniawan, "Pembelajaran Pemrograman Python Dengan Pendekatan Logika Algoritma," *J. Tek. Inform. Mahakarya*, vol. 03, no. 2, pp. 37–44, 2020
- [10] Nasir, N. A., Rasid, N. S., & Ahmad, N. (2014). Grouping Students Academic Performance Using One-Way Clustering. *International journal of Science Commerce and Humanities*, 2(3), 131-138.
- [11] Setiyono, B., & Isnanto, R. R. (2008). Pembagian Kelas Kuliah Mahasiswa Menggunakan Algoritma Pengklasteran Fuzzy C-Means. In *Prosiding Seminar Nasional Teknoin*.
- [12] Li, L., Luo, X., & Chen, H. (2015). Clustering students for group-based learning in foreign language learning. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 9(2), 55-72.
- [13] Syauqoti, R., & Ghozali, M. (2018). Analisis sistem lembaga keuangan syariah Dan lembaga keuangan konvensional. *Jurnal Iqtishaduna*, 14(1), 19-21.