Deteksi Fraud Pada Akun Wifi Universitas Islam Indonesia Dengan Metode Principal Component Analysis

Dio Agus Nofrizal
Program Studi Sarjana Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM. 14.5, Sleman, Yogyakarta, Indonesia
17523110@students.uii.ac.id

Mukhammad Andri Setiawan
Program Studi Sarjana Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM. 14.5, Sleman, Yogyakarta, Indonesia
035230102@uii.ac.id@uii.ac.id

***Abstrak*—Fraud menjadi sebuah masalah yang dapat merugikan orang lain sehingga harus dilakukan tindakan. Fraud terjadi ketika pengguna membagikan akunnya dengan orang lain untuk mengakses wifi UIIConnect tanpa memikirkan celah keamanan yang dapat membahayakan data pengguna. Adanya fraud yang dilakukan dalam penggunaan akun wifi UIIConnect dapat dideteksi dengan menggunakan metode *principal component analysis*. *Principal component analysis* dapat melakukan pengelompokan akun-akun yang terindikasi melakukan fraud atau pun tidak. Tujuan dari penelitian ini adalah untuk mengembangkan model untuk mendeteksi fraud yang akan bermanfaat bagi Badan Sistem Informasi Universitas Islam Indonesia. Dalam proses mencapai tujuan dari penelitian, peneliti menggunakan lima langkah yaitu pengumpulan data, *pre-processing*, *labelling*, *clustering*, dan evaluasi. Dari penelitian ini telah berhasil dikembangkan model yang mampu mendeteksi pengguna yang terindikasi melakukan fraud. Model tersebut berhasil mendapatkan total *variance* sebesar 95,5% dan berhasil mereduksi dimensi komponen. Selanjutnya akun yang terindikasi melakukan fraud akan dilakukan tindakan untuk keamanan data pengguna, sehingga pengguna harus berhati-hati dalam menggunakan akunnya.**

Keywords—fraud, principal component analysis, component, machine learning, clustering.

I. PENDAHULUAN

Universitas Islam Indonesia (UII) merupakan salah satu kampus yang memberikan fasilitas kepada mahasiswa, dosen, dan staf aktif di lingkungan UII untuk dapat mengakses wifi UIIConnect dengan menggunakan akun UII. UIIConnect saat ini telah terpasang lebih dari 700 *Access Points* di seluruh gedung UII. Total *bandwidth* yang disediakan UIIConnect mencapai 3.7 Gbps dan akses per user mencapai 125 Mbps [1]. Adanya fasilitas tersebut beberapa pengguna biasanya membagikan akunnya baik dengan teman atau orang terdekatnya untuk mengakses UIIConnect. Kondisi ini sangat berbahaya karena dapat dideteksi sebagai fraud.

Fraud merupakan penipuan yang dilakukan secara sengaja dengan tujuan untuk mendapatkan keuntungan pribadi yang dapat menyebabkan kerugian bagi orang lain [2]. Orang yang melakukan kejahatan ini biasanya disebut *fraudster*. Dalam kasus ini, ketika fraudster memiliki akun untuk mengakses wifi UIIConnect, akan sangat mungkin bagi *fraudster* untuk dapat mengkases platform lainnya dengan menggunakan akun tersebut. Hal ini dikarenakan seseorang cenderung *login* di berbagai macam platform dengan akun yang sama. Selain itu, *fraudster* yang mendapatkan akses akun orang lain bisa saja akan terjadi penipuan yang berujung pada masalah finansial seperti menargetkan akun bank untuk transfer dana ke akun sendiri atau akun *eCommerce* dan melakukan pembelian palsu.

Banyak aktivitas di kampus UII yang menggunakan wifi UIIConnect untuk mengakses internet setiap harinya. Aktivitas seperti kegiatan belajar mengajar atau aktivitas lain yang dilakukan oleh staf, dosen maupun mahasiswa pasti membutuhkan akses internet dengan menggunakan wifi UII. Terlebih lagi UII memberikan masing-masing akun yang dapat terhubung ke UIIConnect hingga 4 perangkat. Maka dari itu, sangat sulit mengidentifikasi akun yang melakukan fraud karena banyaknya pengguna yang terhubung dengan UIIConnect. Agar dapat mengatasi masalah akun yang melakukan fraud, maka perlu dilakukan analisis setiap lokasi dan akses yang dilakukan ketika menggunakan wifi. Kedua faktor tersebut dapat diketahui dengan melihat *Acces Point* mana pengguna terhubung dan akses yang dilakukan pengguna. Akan tetapi, Badan Sistem Informasi UII mengalami kesulitan untuk mengecek satu persatu akun yang melakukan fraud. Dari permasalahan tersebut, dapat ditarik kesimpulan bahwa Badan Sistem Informasi UII perlu memiliki suatu sistem untuk membantu dalam menganalisis akun yang melakukan fraud. Oleh karena itu, akan dibuat sistem untuk menghitung apakah akun yang menggunakan wifi UIIConnect melakukan fraud atau tidak.

Sistem akan dilengkapi dengan model *principal component analysis* untuk melakukan *clustering*. *Clustering* bertujuan untuk mengelompokkan akun-akun yang terindikasi melakukan fraud atau pun tidak. *Principal component analysis* (PCA) berguna untuk mengurangi dimensi permasalahan menjadi lebih sederhana dengan cara mengidentifikasi sebagian kecil komponen utama dan secara efektif merangkum sebagian besar variasi data [3]. PCA tetap menjaga *variance* sebanyak mungkin agar bisa menemukan variabel baru yang menggambarkan fungsi linier dari kumpulan data asli. Jika terdapat data yang memuat nilai abnormal, karakteristik dari vektor akan sangat berpengaruh karena PCA sangat peka terhadap hal tersebut [4].

Diharapkan sistem yang dihasilkan dapat membantu Badan Sistem Informasi UII untuk mengetahui akun yang terdeteksi melakukan fraud. Selanjutnya Badan Sistem Informasi UII dapat melakukan tindakan terhadap akun yang terdeteksi melakukan fraud. Tindakan tersebut dapat berupa peringatan atau pemblokiran akun.

II. PENELITIAN TERKAIT

Penelitian yang dilakukan sebelumnya memiliki kasus dan cara yang berbeda dalam mendeteksi fraud. Viswanath melakukan penelitian tentang deteksi perilaku anomali pada

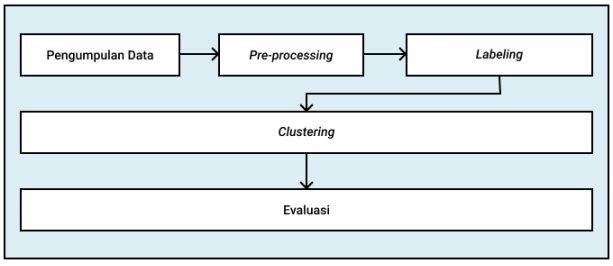
jejaring sosial. Deteksi dilakukan dengan metode *unsupervised learning* yaitu *Principal Component Analisys* (PCA) yang digunakan untuk membedakan perilaku pengguna normal dan tidak normal. PCA memodelkan perilaku pengguna normal secara akurat dan mengidentifikasi anomali secara signifikan. Hasil evaluasi pendekatan yang dilakukan mencapai tingkat deteksi lebih dari 66% dan mencakup lebih dari 94% perilaku buruk dengan positive false kurang dari 0,3% [5]. Selain itu, Meng Bi juga melakukan penelitian tentang *anomaly detection* menggunakan metode PCA. Hasilnya PCA secara akurat dapat menggambarkan perilaku pengguna normal dan anomali serta dapat meningkatkan efisiensi dan stabilitas [4].

Penelitian tentang deteksi anomali juga dilakukan oleh Paul. Penelitian ini menggunakan data jaringan aktivitas pengguna pada organisasi dan perusahaan yang disimpan sebagai *log*. *Log* ini akan digunakan sebagai fitur untuk melatih model dalam melakukan pengelompokan. Penelitian ini melakukan perbandingan metode *Gaussian Mixture Model* (GMM), *K-means* dan *Bayesian Gaussian Mixture Model* (BGMM). GMM menghasilkan *false positive* paling sedikit sebesar 0.33% sedangkan *K-means* 21.77% dan BGMM 5.67% [6].

Terdapat beberapa perbedaan dari penelitian-penelitian sebelumnya. Penelitian ini membahas mengenai anomaly pada penggunaan akun wifi UIIConnect atau terindikasi melakukan fraud. Fraud dapat diketahui berdasarkan faktor lokasi dan akses yang dilakukan pengguna. Sistem akan dilengkapi dengan model *principal component analysis* (PCA) untuk melakukan pengelompokan akun-akun yang terindikasi melakukan fraud atau pun tidak.

III. METODOLOGI PENELITIAN

Bagian ini menjelaskan metodologi penelitian yang digunakan untuk mengidentifikasi pengguna dibagi lima langkah yaitu pengumpulan data, *pre-processing*, *labelling*, *clustering*, dan evaluasi. Gambar 1 menunjukkan metodologi yang digunakan pada penelitian ini.



Gambar 1. Metodologi penelitian

A. Pengumpulan Data

Pada langkah ini, dilakukan pengumpulan data internal yang diambil dari *database* Badan Sistem Informasi UII. Data tersebut berupa file csv yang akan digunakan sebagai data untuk kebutuhan sistem.

B. Pre-processing

Tahap *pre-processing* digunakan untuk mempersiapkan data sebelum dilakukan tahap selanjutnya. Berikut adalah langkah-langkah yang dilakukan pada tahap *pre-processing*:

1) Pengambilan Kolom

Pada tahap ini, akan diambil beberapa kolom yang akan digunakan untuk penelitian. Kolom yang digunakan pada masing-masing data akan mewakili faktor lokasi dan akses yang dilakukan pengguna ketika menggunakan wifi UIIConnect.

2) Mengelompokkan dan menghapus nilai NA

Pada tahap ini, dilakukan pengelompokan aplikasi yang digunakan oleh masing-masing akun. Setelah itu, menghapus seluruh baris yang bernilai *NA*. Hal ini dikarenakan baris tersebut bukan merupakan data akun dan tidak akan digunakan dalam penelitian ini.

3) Memberikan label kemiripan aplikasi

Setelah melakukan pengelompokkan aplikasi, selanjutnya dilakukan pemberian label. Hal ini dilakukan untuk mengetahui kemiripan aplikasi yang dilakukan masing-masing akun.

4) Menghitung persentase kemiripan aplikasi

Selanjutnya persentase label 0 dan 1 pada setiap akun akan dihitung. Tahap ini dilakukan untuk mengetahui persentase kemiripan aplikasi yang digunakan untuk setiap IP yang menggunakan akun yang sama.

5) Menghitung jumlah IP

Pada tahap ini, akan dilakukan perhitungan jumlah IP yang digunakan oleh setiap akun. Dengan mengetahui jumlah IP ini, dapat diketahui jumlah perangkat yang digunakan pengguna ketika menggunakan wifi UIIConnect.

6) Menghitung jumlah Access Point

Pada tahap ini, akan dilakukan perhitungan jumlah *Acces Point* yang digunakan oleh setiap pengguna. Jumlah *access point* tersebut dapat digunakan untuk mengetahui apakah pengguna berada di lokasi yang sama atau tidak.

7) Menghapus karakter domain email

Penghapusan karakter domain email digunakan untuk melakukan *merge* data URL dan data *access point*. Hal ini dikarenakan kedua data memiliki format yang berbeda dalam melakukan *capture* akun pada saat menggunakan wifi UIIConnect.

8) Melakukan merge data

Tahap terakhir adalah melakukan *merge* data URL dan data *access point*. Data tersebut digabungkan berdasarkan akun pada data URL dan data *access point*.

C. Labelling

Pada tahap labelling akan diberikan kondisi untuk memberikan label pada pengguna. Tahap ini digunakan untuk mengetahui pengguna yang terindikasi melakukan fraud.

D. Clustering

Bagian ini menjabarkan tentang proses *clustering* pengguna yang melakukan fraud dengan menggunakan metode *principal component analisys*. Proses ini bertujuan untuk mengelompokkan akun-akun yang terindikasi melakukan fraud atau pun tidak. Berikut adalah tahap-tahap yang dilakukan pada metode PCA:

1) Standardisasi Data

Standardisasi data bertujuan agar setiap variabel memiliki kontribusi yang sama. Berikut adalah hal-hal yang dilakukan pada standarisasi data:

- Menghitung rata-rata menggunakan persamaan:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \quad (1)$$

- Menghitung *center*

$$x^{(i)} = x^{(i)} - \mu \qquad (2)$$

- Menghitung kovarian matrix yang merupakan matrix M x M [7]

$$C = \frac{1}{M} X' X'^{T} \qquad (3)$$

2) Menghitung Variance

Variance meruapakan sebaran data yang ditangkap oleh masing-masing *principal component*. Berikut adalah persamaan untuk menghitung *variance*:

$$\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)\,T} u\right)^2 = \frac{1}{m}\sum_{i=1}^{m} u^T x^{(i)} x^{(i)\,T} u = u^T \left(\sum_{i=1}^{m} x^{(i)} x^{(i)\,T}\right) u \qquad (4)$$

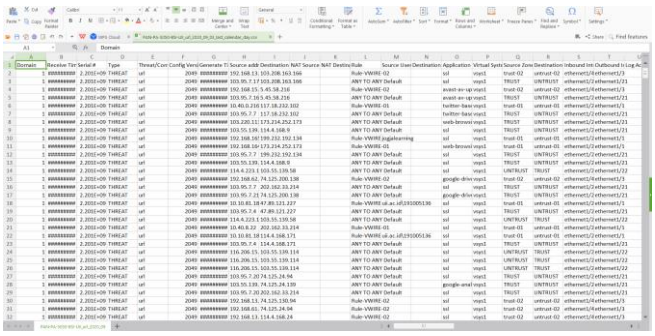
E. Evaluasi

Pada tahap evaluasi akan dilakukan pengukuran performa model. Tahap ini bertujuan untuk mengetahui apakah pemodelan sudah sesuai dengan yang diinginkan serta mengetahui sejauh mana pemodelan ini berhasil.

IV. HASIL

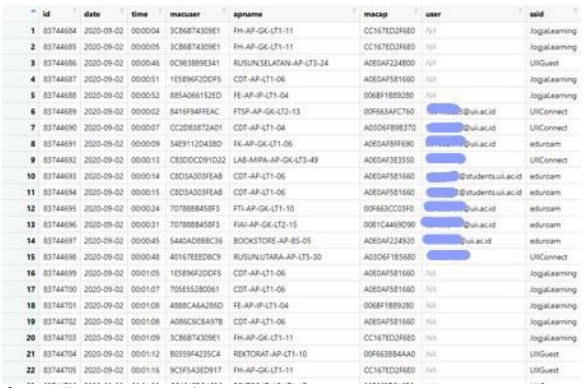
A. Pengumpulan Data

Data yang dibutuhkan dalam pembuatan model ini berupa hasil *record* setiap akun yang menggunakan wifi UII. Terdapat dua data yang digunakan pada penelitian yaitu data URL dan data *access point*. Gambar 3 memuat data URL yang berisi *receive time*, *source user*, *source address*, *application* dan lain-lain.



Gambar 3. Data URL

Gambar 4 memuat data *access point* yang berisi *id*, *date*, *time*, *mac user* dan lain-lain.



	id	date	time	macuser	apname	macap	user	ssid
1	83744684	2020-09-02	00:00:04	3CB687430	FH-AP-GK-LT1-11			JogjaLearning
2	83744685	2020-09-02	00:00:05	3CB687430	FH-AP-GK-LT1-11			JogjaLearning
3	83744686	2020-09-02	00:00:46	0C983889E	RUSUN.SELATAN-AP-LT3-24			UIIGuest
4	83744687	2020-09-02	00:00:51	1E5B96F2	CDT-AP-LT1-06			JogjaLearning
5	83744688	2020-09-02	00:00:52	885A06615	FE-AP-IP-LT1-04			JogjaLearning
6	83744689	2020-09-02	00:00:02	8416F94FF	FTSP-AP-GK-LT2-13			UIIConnect
7	83744690	2020-09-02	00:00:07	CC2D83872	CDT-AP-LT1-04			UIIConnect
8	83744691	2020-09-02	00:00:09	94E9112D4	FK-AP-GK-LT1-06			eduroam
9	83744692	2020-09-02	00:00:13	C83DDCD91	LAB-MIPA-AP-GK-LT3-49			UIIConnect
10	83744693	2020-09-02	00:00:14	C8D3A303F	CDT-AP-LT1-06			eduroam
11	83744694	2020-09-02	00:00:15	C8D3A303F	CDT-AP-LT1-06			eduroam
12	83744695	2020-09-02	00:00:24	7078BB458	FTI-AP-GK-LT1-10			eduroam
13	83744696	2020-09-02	00:00:31	7078BB458	FIAI-AP-GK-LT2-15			eduroam
14	83744697	2020-09-02	00:00:45	5440AD88B	BOOKSTORE-AP-BS-05			eduroam
15	83744698	2020-09-02	00:00:48	40167EED8	RUSUN.UTARA-AP-LT5-30			UIIConnect
16	83744699	2020-09-02	00:01:05	1E5B96F2	CDT-AP-LT1-06			JogjaLearning
17	83744700	2020-09-02	00:01:07	705E5528	CDT-AP-LT1-06			JogjaLearning
18	83744701	2020-09-02	00:01:08	4888CA6A2	FE-AP-IP-LT1-04			JogjaLearning
19	83744702	2020-09-02	00:01:08	A086C6C8A	CDT-AP-LT1-06			JogjaLearning
20	83744703	2020-09-02	00:01:09	3CB687430	FH-AP-GK-LT1-11			JogjaLearning
21	83744704	2020-09-02	00:01:12	B0359F423	REKTORAT-AP-LT1-10			UIIGuest
22	83744705	2020-09-02	00:01:16	9C5F5A3ED	FH-AP-GK-LT1-11			UIIGuest

Gambar 4. Data *access point*

B. Pre-processing

Setelah melakukan pengumpulan data, selanjutnya dilakukan pre-processing data. Berikut adalah langkah-langkah yang dilakukan pada tahap *pre-processing*:

1) Pengambilan Kolom

Pada tahap ini, diambil beberapa kolom dari masing-masing data yang akan digunakan pada penelitian. Pada data URL kolom yang diambil yaitu "Source.User", "Source.address" dan "Aplikasi". Kolom tersebut digunakan untuk mengetahui akun yang digunakan, jumlah perangkat dan aplikasi apa saja yang diakses. Gambar 5 merupakan hasil dari pengambilan kolom data URL.

	Source.User	Source.address	Application
1	NA	192.168.13.15	ssl
2	NA	103.95.7.17	ssl
3	NA	192.168.15.11	avast-av-update
4	NA	103.95.7.16	avast-av-update
5	NA	10.40.0.216	twitter-base
6	NA	103.95.7.7	twitter-base
7	NA	103.220.113.12	web-browsing
8	NA	103.55.139.35	ssl
9	jogjalearning	192.168.165.109	ssl
10	NA	192.168.164.254	web-browsing
11	NA	103.95.7.7	ssl
12	NA	103.55.139.35	ssl
13	NA	114.4.223.140	ssl
14	NA	192.168.62.89	google-drive-web
15	NA	103.95.7.7	ssl
16	NA	103.95.7.21	google-drive-web
17	uii.ac.id\191	10.10.81.18	ssl
18	NA	103.95.7.4	ssl

Showing 1 to 19 of 800,000 entries. 3 total columns

Gambar 5. Hasil pengambilan kolom data url

Selanjutnya pada data *access point* kolom yang diambil hanya dua yaitu 'user' dan 'apname'. Kolom ini digunakan untuk mengetahui lokasi pengguna ketika menggunakan wifi UIIConnect. Gambar 6 merupakan hasil dari pengambilan kolom data *access point*.

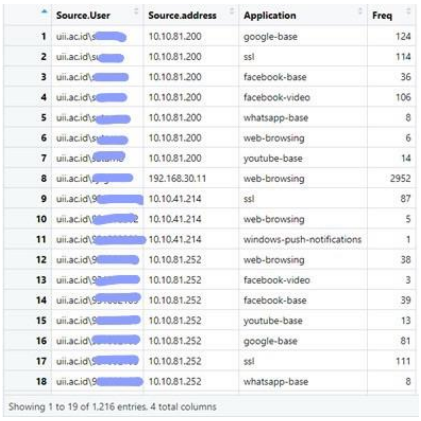
	user	apname
1	NA	FH-AP-GK-LT1-11
2	NA	FH-AP-GK-LT1-11
3	NA	RUSUN.SELATAN-AP-LT3-24
4	NA	CDT-AP-LT1-06
5	NA	FE-AP-IP-LT1-04
6	@uii.ac.id	FTSP-AP-GK-LT2-13
7	@uii.ac.id	CDT-AP-LT1-04
8	@uii.ac.id	FK-AP-GK-LT1-06
9		LAB-MIPA-AP-GK-LT3-49
10	@students.uii.ac.id	CDT-AP-LT1-06
11	@students.uii.ac.id	CDT-AP-LT1-06

Gambar 6. Hasil pengambilan kolom data *access point*

2) Mengelompokkan dan menghapus nilai NA

Tahap selanjutnya adalah mengelompokkan aplikasi yang digunakan oleh masing-masing akun dan mengahapus baris yang bernilai *NA*. Gambar 7 merupakan hasil setelah

mengelompokkan aplikasi dan menghapus nilai *NA* pada data URL.

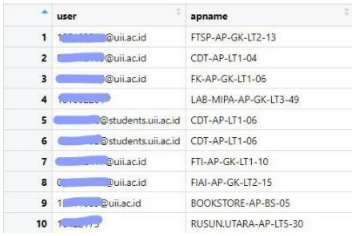


	Source.User	Source.address	Application	Freq
1	uii.ac.id\s[redacted]	10.10.81.200	google-base	124
2	uii.ac.id\s[redacted]	10.10.81.200	ssl	114
3	uii.ac.id\s[redacted]	10.10.81.200	facebook-base	36
4	uii.ac.id\s[redacted]	10.10.81.200	facebook-video	106
5	uii.ac.id\s[redacted]	10.10.81.200	whatsapp-base	8
6	uii.ac.id\s[redacted]	10.10.81.200	web-browsing	6
7	uii.ac.id\s[redacted]	10.10.81.200	youtube-base	14
8	uii.ac.id\[redacted]	192.168.30.11	web-browsing	2952
9	uii.ac.id\9[redacted]	10.10.41.214	ssl	87
10	uii.ac.id\9[redacted]	10.10.41.214	web-browsing	5
11	uii.ac.id\9[redacted]	10.10.41.214	windows-push-notifications	1
12	uii.ac.id\9[redacted]	10.10.81.252	web-browsing	38
13	uii.ac.id\9[redacted]	10.10.81.252	facebook-video	3
14	uii.ac.id\9[redacted]	10.10.81.252	facebook-base	39
15	uii.ac.id\9[redacted]	10.10.81.252	youtube-base	13
16	uii.ac.id\9[redacted]	10.10.81.252	google-base	81
17	uii.ac.id\9[redacted]	10.10.81.252	ssl	111
18	uii.ac.id\9[redacted]	10.10.81.252	whatsapp-base	8

Showing 1 to 19 of 1,216 entries, 4 total columns

Gambar 7. Hasil mengelompokkan aplikasi dan menghapus nilai *NA*

Sedangkan pada data *access point* hanya menghapus baris yang bernilai *NA*. Gambar 8 merupakan hasil penghapusan nilai *NA* pada data *access point.*

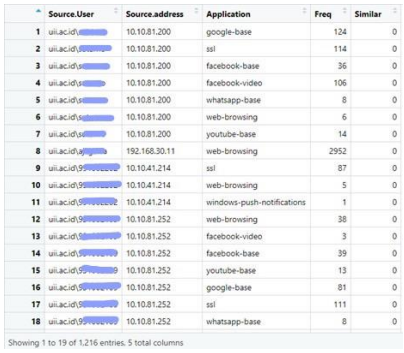


	user	apname
1	1[redacted]@uii.ac.id	FTSP-AP-GK-LT2-13
2	[redacted]@uii.ac.id	CDT-AP-LT1-04
3	[redacted]@uii.ac.id	FK-AP-GK-LT1-06
4	[redacted]	LAB-MIPA-AP-GK-LT3-49
5	[redacted]@students.uii.ac.id	CDT-AP-LT1-06
6	[redacted]@students.uii.ac.id	CDT-AP-LT1-06
7	[redacted]@uii.ac.id	FTI-AP-GK-LT1-10
8	0[redacted]@uii.ac.id	FIAI-AP-GK-LT2-15
9	1[redacted]@uii.ac.id	BOOKSTORE-AP-BS-05
10	1[redacted]	RUSUN.UTARA-AP-LT5-30

Gambar 8. Hasil penghapusan nilai *NA*

3) Memberikan label kemiripan apliaksi

Selanjutnya setiap akun akan diberikan label pada kolom ‘Similar’. Akun akan diberikan label 1 jika terdapat kesamaan aplikasi dan label 0 jika terdapat perbedaan aplikasi yang digunakan oleh masing-masing IP pada akun yang sama. Pemberian label tersebut hanya dilakukan pada data URL. Gambar 9 merupakan hasil dari pemberian label pada kolom ‘similar’.



	Source.User	Source.address	Application	Freq	Similar
1	uii.ac.id\s[redacted]	10.10.81.200	google-base	124	0
2	uii.ac.id\s[redacted]	10.10.81.200	ssl	114	0
3	uii.ac.id\s[redacted]	10.10.81.200	facebook-base	36	0
4	uii.ac.id\s[redacted]	10.10.81.200	facebook-video	106	0
5	uii.ac.id\s[redacted]	10.10.81.200	whatsapp-base	8	0
6	uii.ac.id\s[redacted]	10.10.81.200	web-browsing	6	0
7	uii.ac.id\s[redacted]	10.10.81.200	youtube-base	14	0
8	uii.ac.id\[redacted]	192.168.30.11	web-browsing	2952	0
9	uii.ac.id\9[redacted]	10.10.41.214	ssl	87	0
10	uii.ac.id\9[redacted]	10.10.41.214	web-browsing	5	0
11	uii.ac.id\9[redacted]	10.10.41.214	windows-push-notifications	1	0
12	uii.ac.id\9[redacted]	10.10.81.252	web-browsing	38	0
13	uii.ac.id\9[redacted]	10.10.81.252	facebook-video	3	0
14	uii.ac.id\9[redacted]	10.10.81.252	facebook-base	39	0
15	uii.ac.id\9[redacted]	10.10.81.252	youtube-base	13	0
16	uii.ac.id\9[redacted]	10.10.81.252	google-base	81	0
17	uii.ac.id\9[redacted]	10.10.81.252	ssl	111	0
18	uii.ac.id\9[redacted]	10.10.81.252	whatsapp-base	8	0

Showing 1 to 19 of 1,216 entries, 5 total columns

Gambar 9. Hasil pemberian label

4) Menghitung persentase kemiripan aplikasi

Setelah memberikan label, selanjutnya akan dihitung persentase label 0 yang berada pada kolom ‘Similar0’ dan persentase label 1 pada kolom ‘Similar1’ untuk setiap akun. Kolom ‘Similar’ merupakan persentase ketidakmiripan aplikasi yang digunakan sedangkan kolom “Similar1” merupakan persentase kemiripan aplikasi yang digunakan. Gambar 10 merupakan hasil dari perhitungan persentase.

	Source.User	Similar0	Similar1
1	0[redacted]	100.000000	0.00000
2	[redacted]@uii.ac.id	100.000000	0.00000
3	1[redacted]	14.285714	85.71429
4	1[redacted]@uii.ac.id	100.000000	0.00000
5	[redacted]	100.000000	0.00000
6	[redacted]	100.000000	0.00000
7	1[redacted]	100.000000	0.00000
8	1[redacted]	100.000000	0.00000
9	1[redacted]	100.000000	0.00000
10	1[redacted]	100.000000	0.00000
11	[redacted]	100.000000	0.00000
12	1[redacted]	100.000000	0.00000
13	9[redacted]	100.000000	0.00000
14	[redacted]	71.428571	28.57143
15	[redacted]@mail.umy.ac.id	100.000000	0.00000
16	[redacted]	6.046512	93.95349
17	[redacted]	100.000000	0.00000
18	[redacted]	100.000000	0.00000

Showing 1 to 19 of 143 entries, 3 total columns

Gambar 10. Hasil perhitungan persentase

5) Menghitung jumlah IP

Pada tahap ini, akan dilakukan perhitungan jumlah IP yang digunakan oleh setiap pengguna. Gambar 11 merupakan hasil dari perhitungan jumlah IP.

	Source.User	freq
1	0[redacted]	1
2	1[redacted]@uii.ac.id	1
3	1[redacted]	2
4	1[redacted]@uii.ac.id	1
5	1[redacted]	1
6	1[redacted]	1
7	1[redacted]	1
8	1[redacted]	1
9	1[redacted]	1
10	1[redacted]	1
11	1[redacted]	1
12	1[redacted]	1
13	9[redacted]	1
14	e[redacted]	2
15	f[redacted]@mail.umy.ac.id	1
16	[redacted]	23
17	[redacted]	1
18	k[redacted]	1

Showing 1 to 19 of 143 entries, 2 total columns

Gambar 11. Hasil perhitungan ip

6) Menghitung jumlah Access Point

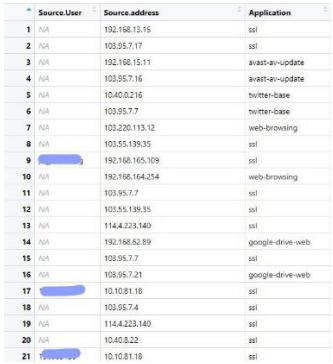
Pada tahap ini, akan dilakukan perhitungan jumlah *acces point* yang digunakan oleh setiap akun. Gambar 12 merupakan hasil dari perhitungan jumlah *access point* yang digunakan oleh setiap akun.

	user	freq
1	0[redacted]	1
2	[redacted]	2
3	[redacted]	1
4	[redacted]	1
5	[redacted]	1
6	[redacted]	1
7	[redacted]	1
8	[redacted]	1
9	[redacted]	2
10	[redacted]	2
11	[redacted]	2
12	[redacted]	3

Gambar 12. Hasil perhitungan jumlah *access*

7) Menghapus karakter domain email

Selanjutnya menghapus karakter domain email pada data URL dan data *access point*. Pada data URL, karakter yang dihapus adalah karakter uii\ac\id\ dan karakter setelah @ seperti @uii.ac.id. Gambar 13 merupakan hasil dari penghapusan karakter domain email pada data URL.



	Source.User	Source.address	Application
1	NA	192.168.13.15	ssl
2	NA	103.95.7.17	ssl
3	NA	192.168.15.11	avast-av-update
4	NA	103.95.7.16	avast-av-update
5	NA	10.40.0.216	twitter-base
6	NA	103.95.7.7	twitter-base
7	NA	103.220.113.12	web-browsing
8	NA	103.55.139.35	ssl
9	[redacted]	192.168.165.109	ssl
10	NA	192.168.164.254	web-browsing
11	NA	103.95.7.7	ssl
12	NA	103.55.139.35	ssl
13	NA	114.4.223.140	ssl
14	NA	192.168.62.89	google-drive-web
15	NA	103.95.7.7	ssl
16	NA	103.95.7.21	google-drive-web
17	[redacted]	10.10.81.18	ssl
18	NA	103.95.7.4	ssl
19	NA	114.4.223.140	ssl
20	NA	10.40.8.22	ssl
21	[redacted]	10.10.81.18	ssl

Gambar 13. Hasil penghapusan karakter domain email pada data url

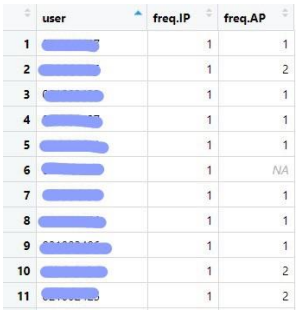
Sedangkan pada data *access point* penghapusan karakter domain email hanya dilakukan pada karakter setelah @ seperti @students.ui.ac.id. Gambar 14 merupakan hasil penghapusan karakter pada data access point.

	user	apname
1	[redacted]	FTSP-AP-GK-LT2-13
2	[redacted]	CDT-AP-LT1-04
3	[redacted]	FK-AP-GK-LT1-06
4	[redacted]	LAB-MIPA-AP-GK-LT3-49
5	[redacted]	CDT-AP-LT1-06
6	[redacted]	FTI-AP-GK-LT1-10
7	[redacted]	FIAI-AP-GK-LT2-15
8	[redacted]	BOOKSTORE-AP-BS-05
9	[redacted]	RUSUN.UTARA-AP-LT5-30

Gambar 14. Hasil penghapusan karakter domain email pada data *access point*

8) Melakukan merge data

Tahap terakhir adalah melakukan *merge* jumlah IP pada data URL dan jumlah AP pada data *access point*. Data tersebut digabungkan berdasarkan kolom 'Source.address' pada data URL. Gambar 15 merupakan hasil penggabungan jumlah IP pada data URL dan jumlah AP pada data *access point*.



	user	freq.IP	freq.AP
1	[redacted]	1	1
2	[redacted]	1	2
3	[redacted]	1	1
4	[redacted]	1	1
5	[redacted]	1	1
6	[redacted]	1	NA
7	[redacted]	1	1
8	[redacted]	1	1
9	[redacted]	1	1
10	[redacted]	1	2
11	[redacted]	1	2

Gambar 15. Hasil penggabungan data URL dan data *access point*

Gambar 16 merupakan hasil akhir dari proses *pre-processing* yang menyajikan kolom 'akun', 'JumlahIP', 'JumlahAP' dan kolom 'PersentaseKetidakmiripan'. Kolom 'JumlahIP' dan 'JumlahAP' dapat menggambarkan lokasi pengguna sedangkan kolom "PersentaseKetidakmiripan" dapat menggambarkan akses yang dilakukan pada saat menggunakan wifi UIIConnect.

	JumlahIP	JumlahAP	PersentaseKetidakmiripan
[redacted]	1	1	100.000000
[redacted]	1	2	100.000000
[redacted]	1	1	100.000000
[redacted]	1	1	100.000000
[redacted]	1	1	100.000000
[redacted]	1	NA	100.000000
[redacted]	1	1	100.000000
[redacted]	1	1	100.000000
[redacted]	1	1	100.000000
[redacted]	1	2	100.000000
[redacted]	1	2	100.000000
[redacted]	1	2	100.000000
[redacted]	1	3	100.000000
[redacted]	1	1	100.000000
[redacted]	1	5	100.000000
[redacted]	1	1	100.000000
[redacted]	1	1	100.000000

Gambar 16. Hasil p*re-processing*

C. Labelling

Dalam tahap *labelling*, masing-masing pengguna akan diberikan label antara 1 atau 0. Label 1 menunjukkan jika pengguna memenuhi kondisi 'JumlahIP' dan 'JumlahAP' lebih dari satu serta kolom 'PersentaseKetidakmiripan' lebih dari 50, jika kondisi sebaliknya maka diberikan label 0. Gambar 17 menunjukkan tabel hasil *labelling* masing-masing pengguna.

	user	JumlahIP	JumlahAP	PersentaseKetidakmiripan	Label
37	[redacted]	2	4	66.666667	1
57	[redacted]	2	2	75.000000	1
89	[redacted]	2	2	71.428571	1
91	[redacted]	2	4	72.413793	1
110	[redacted]	2	2	66.666667	1
1	[redacted]	1	1	100.000000	0
2	[redacted]	1	2	100.000000	0
3	[redacted]	1	1	100.000000	0
4	[redacted]	1	1	100.000000	0
5	[redacted]	1	1	100.000000	0
6	[redacted]	1	NA	100.000000	0
7	[redacted]	1	1	100.000000	0
8	[redacted]	1	1	100.000000	0
9	[redacted]	1	1	100.000000	0

Gambar 17. Hasil pemberian label

D. Clustering

Pada tahap *clustering* akan dilakukan tahap standardisasi data dan menghitung *variance*.

1) Standardisasi Data

Standardisasi data dilakukan dengan menggunakan skala sehingga data akan memiliki *impact* yang sama dan *comparable.*

a) Menghitung rata-rata

Gambar 18 merupakan hasil perhitungan mean masing-masing variabel. Variabel tersebut berupa 'JumlahIP' (1,167), 'JumlahAP' (1,745) dan 'PersentaseKetidakmiripan' (94,00).

```
> summary(training)
    JumlahIP        JumlahAP      PersentaseKetidakmiripan    Label
Min.   :1.000   Min.   : 1.000   Min.   : 33.33           Length:102
1st Qu.:1.000   1st Qu.: 1.000   1st Qu.:100.00           Class :character
Median :1.000   Median : 1.000   Median :100.00           Mode  :character
Mean   :1.167   Mean   : 1.745   Mean   : 94.00
3rd Qu.:1.000   3rd Qu.: 2.000   3rd Qu.:100.00
Max.   :3.000   Max.   :16.000   Max.   :100.00
```

Gambar 18. Hasil perhitungan rata-rata

b) Menghitung center

Selanjutnya pada Gambar 19 akan dilakukan *centering* untuk setiap variabel. Variabel 'JumlahAP' menghasilkan nilai *centering* (1,66667), 'JumlahAP' (1,745098) dan 'PersentaseKetidakmiripan' (94,001822).

```
> pc$center
                JumlahIP                 JumlahAP PersentaseKetidakmiripan
                1.166667                 1.745098                94.001822
```

Gambar 19. Hasil perhitungan center

c) Menghitung kovarian matrix

Pada gambar 20 merupakan hasil perhitungan kovarian matrix. Dari perhitungan tersebut, dapat dilihat kekuatan korelasi masing-masing variabel dengan setiap *principal component*.

```
> pc$rotation
                                PC1         PC2         PC3
JumlahIP                0.70636196 -0.03517388  0.706976359
JumlahAP               -0.04234469 -0.99907567 -0.007398678
PersentaseKetidakmiripan -0.70658312  0.02471055  0.707198478
```

Gambar 20. Hasil perhitungan kovarian matrix

Gambar 21 merupakan hasil akhir dalam tahap standardisasi data.

```
> pc$x
                    PC1         PC2         PC3
001002407 -0.5232992  0.4339987 -0.01142491
001002425 -0.5466769 -0.1175715 -0.01550958
001002433 -0.5232992  0.4339987 -0.01142491
001002437 -0.5232992  0.4339987 -0.01142491
011002421 -0.5232992  0.4339987 -0.01142491
011002428 -0.5232992  0.4339987 -0.01142491
011002444 -0.5232992  0.4339987 -0.01142491
021002406 -0.5232992  0.4339987 -0.01142491
021002408 -0.5466769 -0.1175715 -0.01550958
021002425 -0.5466769 -0.1175715 -0.01550958
021002429 -0.5466769 -0.1175715 -0.01550958
```

Gambar 21. Hasil standardisasi data

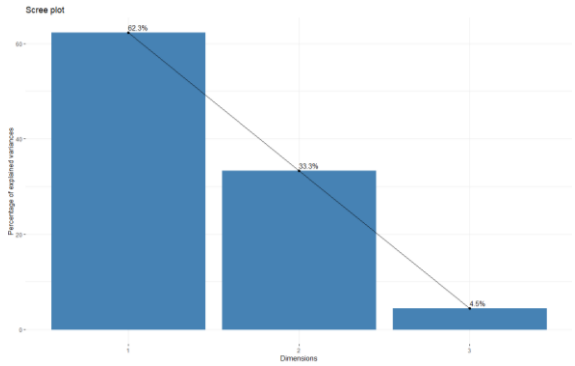
2) Menghitung Variance

Gambar 22 merupakan hasil dari perhitungan *variance* setiap *principal component*. Principal component (PC) merupakan dimensi baru yang merangkum banyaknya informasi yang ada di seluruh variabel sebelumnya. Hasil *variance* yang didapat oleh 'PC1' (0.6226), 'PC2' (0.3328) dan 'PC3' (0,04454).

```
> summary(pc)
Importance of components:
                          PC1    PC2     PC3
Standard deviation     1.3667 0.9992 0.36552
Proportion of Variance 0.6226 0.3328 0.04454
Cumulative Proportion  0.6226 0.9555 1.00000
```

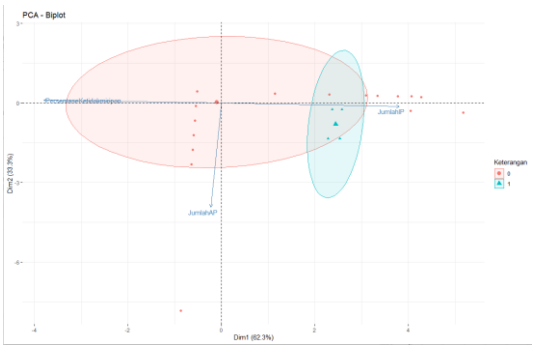
Gambar 22. Hasil Perhitungan *Variance*

Selanjutnya *variance* tersebut dapat dilakukan ploting untk melihat total *variance* yang dibawa oleh masing-masing *component*.



Gambar 23. Ploting *variance*

Pada gambar 23 PC1 memiliki *variance* paling besar dengan nilai 62,3% dan PC3 dengan nilai terkecil 4,5%. Dari ploting tersebut dengan mengambil komponen PC1 dan PC2 maka akan didapatkan total *variance* dari seluruh data sebesar 95,5 % sehingga proses komputasi menjadi lebih baik karena tidak melakukan komputasi semua komponen.



Gambar 24. Hasil Biplot PC1 dan PC2

Selanjutnya pada gambar 18 dapat dilihat hasil biplot pada sumbu vertikal (Dim1) dan sumbu horizontal (Dim2). Kemudian pada biplot di atas juga divisualisasikan setiap variabel sebagai bentuk vektor. Dari plot tersebut, dapat dilihat arah vector 'JumlahIP' dan 'PersentaseKetidakmiripan' cenderung horizontal seperti arah *principal component* yang kedua (Dim2). Hal ini mengindikasikan bahwa variabel 'JumlahIP' dan 'PersentaseKetidakmiripan' lebih banyak dijelaskan atau diwakili oleh *principal component* yang kedua. Sebaliknya, arah vektor 'JumlahAP' lebih mendekati arah *principal component* yang pertama (Dim1). Hal ini mengindikasikan jika informasi yang dibawa variabel 'JumlahAP' lebih banyak diwakili oleh *principal component* yang pertama.

E. Evaluasi

Setelah mendapatkan *cluster* dalam pemodelan, selanjutnya dilakukan evaluasi untuk melihat apakah sudah sesuai dengan yang diinginkan. Tetapi pada tahap ini belum dapat dilakukan karena belum bisa ditentukan evaluasi yang sesuai untuk sistem.

V. KESIMPULAN

Berdasarkan hasil dari pembuatan model yang dilakukan dalam penelitian ini, dapat disimpulkan bahwa penelitian ini telah berhasil melakukan *clustering* penggguna yang melaukan fraud dan tidak. PCA berhasil mendapatkan total *variance* sebesar 95,5% yang hanya menggunakan dua *component principal*. PCA berhasil mereduksi dimensi komponen sehingga proses komputasi menjadi lebih baik karena tidak melakukan komputasi semua komponen. Serta mendapatkan hasil *clustering* pengguna yang terindikasi melakukan fraud dan tidak.

REFERENSI

[1] "Akses Internet," 2017. [Online]. Available: https://bsi.uii.ac.id/akses-internet/.

[2] Z. Zojaji, R. E. Atani, and A. H. Monadjemi, "A Survey of Credit Card Fraud Detection Techniques : Data and Technique Oriented Perspective," pp. 1–26, 2016.

[3] Y. Aït-sahalia and D. Xiu, "Using principal component analysis to estimate a high dimensional factor model with high-frequency data ☆," vol. 201, pp. 384–399, 2017.

[4] M. Bi, J. Xu, M. Wang, and F. Zhou, "Anomaly

detection model of user behavior based on principal component analysis," *J. Ambient Intell. Humaniz. Comput.*, 2016.

[5] B. Viswanath *et al.*, "Towards Detecting Anomalous User Behavior in Online Social Networks," 2014.

[6] M. Paul and K. Medhe, "Using Machine Learning to Detect Anomalies in Internet Browsing Pattern of Users," 2019.

[7] P. N. Primandari and B. Hardiansyah, "Ekstraksi Fitur Menggunakan Principal Component Analisys (PCA)," 2018.