OPEN ACCESS
CENTIVE 2019

ISBN : 978-602-53004-0-0

Conference on Electrical Engineering, Telematics, Industrial Technology, and Creative Media 2019

Deteksi *Fraud* Menggunakan Metode *K-Means* dan *Euclidean Distance* dalam Sensor IoT

Muhammad Idham Habibie

Teknik Elektro, Teknik, Universitas Indonesia
Depok 16951, Indonesia

muhammad.idham.habibie@gmail.com

Abstrak

Dalam era industri 4.0 yang memperkenalkan dunia komunikasi antar D2D membuat *hackers* menjadi lebih memiliki kesempatan untuk bisa mengakses sebuah informasi di Internet. Beberapa metode dari *machine learning* telah dilakukan untuk bisa mendeteksi *fraud* dengan berbagai macam algoritma. Di dalam penelitian ini, penulis menggunakan metode *K-Means* dan *Euclidean Distance* untuk mengukur dan mengidentifikasi *Fraud* dalam sebuah perangkat Raspberry Pi-3 dengan tambahan sensor yaitu *Ultrasonic Sensor* dan *DHT11*. Tujuan dari penggunaan metode ini adalah mengembangkan sebuah metode *machine learning* yang sudah ada untuk menjadi lebih efektif dari sebuah dataset sensor IoT yang jumlahnya relatif besar. Dataset yang diambil adalah 4 variabel diantaranya jarak, kelembapan, suhu, dan *timestamp* selama 1 minggu dengan total dataset sebanyak 9981. Hasil dari metode ini di bandingkan dengan metode *Outlier Detection*, dengan menggunakan *confusion matrix*. Akurasi hasil dari metode ini adalah 99.7% yang relatif baik untuk digunakan dalam deteksi anomali data.

Kata Kunci: *Confusion Matriks, Euclidean Distance, K-Means, Outlier Detection, Raspberry Pi*

I. PENDAHULUAN

ERA industri 4.0 di dunia membuat disrupsi yang sangat pesat terutama pada trafik *mobile* dan trafik *fixed* di dalam sebuah jaringan. Era ini memperkenalkan sebuah interaksi antar perangkat *Device-to-Device* (*D2D*) yang menyebabkan banyaknya kemungkinan adanya *fraud* karena banyaknya trafik yang masuk ke dalam sebuah jaringan. Selain itu, dengan adanya koneksi *Internet Protocol* (IP) layer, *hacker* akan relatif lebih mudah untuk memberikan data *fraud* ketika *surfing* ke dalam jaringan Internet. Apalagi jika jaringan IP ini bisa diakses secara publik, tentunya potensi sebuah devais yang berbasis IP ini akan relatif mudah terkena *fraud.*

Dengan adanya trafik yang berasal dari perangkat *hardware* seperti *handphone*, *desktop*, atau bahkan perangkat *Internet of Things* (IoT), secara tidak sadar akan meningkatkan total jumlah trafik data yang cukup signifikan ke dalam sebuah jaringan *network* yang berbasis IP. Salah satu kasus pernah terjadi dalam sebuah transaksi kartu kredit yang memanfaatkan jaringan data Internet IP, di mana fraud transaksi palsu teridentifikasi sekitar 10 – 20% dari total transaksi *credit card* dalam 1 tahun tersebut.

Dalam paper [1], metode deteksi *fraud* yang digunakan dalam transaksi kartu kredit palsu adalah *K-Nearest Neighbor* (KNN) dan *Outlier Detection* dengan hasil yang relatif sangat baik untuk deteksi *fraud*. Namun, kerugian menggunakan metode KNN ini adalah *limitation memory* [1], yang menyebabkan kurang cocok digunakan dalam dataset yang relatif besar. Dalam paper ini hasil dari *KNN* terbukti bisa meningkatkan deteksi

rasio *fraud* untuk mengkategorikan sebuah anomali dari transaksi yang normal. Selain itu, di *case* yang sama, sebuah penelitian telah mencoba untuk membandingkan performansi setiap klasifikasi seperti *Random Forest*, *Decision Trees*, KNN. Menurut riset ini, seluruh metode klasifikasi memiliki nilai *output confusion matriks* yang relatif sama, namun *Random Forest* adalah salah satu metode klasifikasi yang relatif menonjol dibandingkan dengan yang lain [2].

Di dunia telekomunikasi, penelitian [3] melakukan deteksi anomali dari sebuah total panggilan dalam sistem *billing* telekomunikasi. Untuk mendeteksi panggilan telekomunikasi ini, *Naïve Bayes* sebagai salah satu metode klasifikasi *fraud* digunakan sebagai penelitian untuk mendeteksi anomali tersebut [3]. Metode tersebut dalam penelitian ini bisa cukup membuktikan panggilan yang normal dan *fraud* dalam sebuah skala probabilitas, dan terbukti sesuai dengan target penelitian dalam paper tersebut *fraud* dalam sebuah komunikasi yang berbasis IP, terjadi karena 2 hal, yaitu *Private Branch Exchange* (PBX) *hacking* dan perangkat *handset* VoIP *hacking*. Kedua probabilitas *Fraud* menyebabkan terjadinya sebuah *Call Forwarding Fraud*, *Payphone Fraud* (telepon gratis dalam sebuah operator).

Beberapa kasus dalam teknologi telekomunikasi via IP memotivasi peneliti untuk bisa mengkaji lebih lanjut bagaimana sistem *fraud* ini bekerja dalam IP basis. Tentunya adanya isu terkait *security* yang dikhawatirkan jika *port* IP tersebut dibuka secara publik, tanpa menggunakan *Secure Socket Layer* (SSL), membuat kemungkinan *fraud* menjadi lebih rentan, dan para *hacker* relatif lebih mudah mengakses IP ini. Penelitian ini mencoba membuka sebuah IP Publik yang dapat di akses via *Network Address Translation* (NAT) dimana probabilitas terkena pendayaan data jauh lebih mudah. Oleh karenanya, tujuan dari penelitian ini adalah mengembangkan sebuah algoritma yang efektif untuk bisa mendeteksi anomali sebuah dataset, yang implementasinya dapat digunakan untuk aplikasi *Internet of Things* (IoT) dengan jumlah sensor yang relatif besar.

Dalam penelitian ini, peneliti menggunakan metode *unsupervised learning*, yaitu *K-Means* sebagai salah satu cara untuk mendeteksi kecurangan data. Pertimbangan untuk mengambil *K-Means* adalah metode ini dapat digunakan untuk dataset yang variatif, dengan total dataset yang relatif sedikit dan besar. Dibandingkan dengan metode *clustering* lainnya seperti *Hierarichal Clustering*, metode ini tidak cocok untuk dataset yang relatif besar [4].

Selain itu, meskipun dalam studi literatur untuk identifikasi kecurangan data dalam Kartu Kredit yang dipelajari menggunakan *K-Nearest Neighbor* (KNN) dengan metode *supervised learning*, hal ini menginspirasi peneliti untuk bisa menggunakan metode yang berdekatan seperti *K-Means*. *K-Means* adalah salah satu metode yang relatif mudah digunakan dalam data sains, untuk mengklasifikasi sebuah dataset.

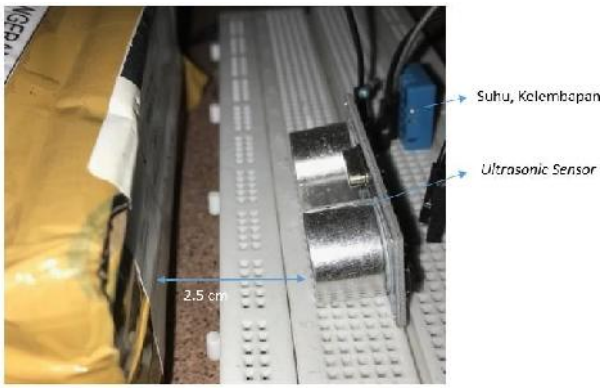
Sebuah pengembangan metode yang ada dengan metode baru akan dibahas dalam riset ini. Riset ini dibagi menjadi beberapa Bab. Bab 1 membahas latar belakang terjadinya sebuah *fraud* dalam sensor IoT yang sudah berbasis IP. Selanjutnya, Bab 2 membahas detail metode penelitian yang digunakan dalam penelitian ini, termasuk cara pengambilan dataset dan pengolahan datanya. Hasil penelitian dari dataset akan dibahas dalam Bab 3, yang merupakan Bab pembahasan bahasa pemrograman. Bab 4 membahas evaluasi dari hasil tersebut, dan komparasi dengan metode lainnya, lalu ditutup oleh Bab 5 sebagai kesimpulan dari penelitian ini.

II. METODE PENELITIAN

A. *Pengambilan Dataset*

Penelitian ini melakukan sebuah simulasi dataset yang berjalan selama 1 minggu, dengan menggunakan sebuah perangkat Mini-PC (*Raspberry Pi*) yang dapat diakses via *Port Forwarding* dalam sebuah router tertentu. Peneliti mengambil sebuah sampel dari sebuah dataset yang didapatkan dari *Ultrasonic Sensor + DHT 11 Sensor* untuk menentukan jarak, kelembapan, suhu, dan *timestamp*.

Peneliti melakukan sebuah simulasi untuk mengukur sebuah benda di depan sensor dengan yang relatif dekat jarak 2,5 cm. Selain itu, simulasi ini juga mengambil data suhu dan kelembapan dari sensor *DHT11* dan mengaitkannya dengan jarak ini.



Gambar 1. Jarak antara sensor dengan benda

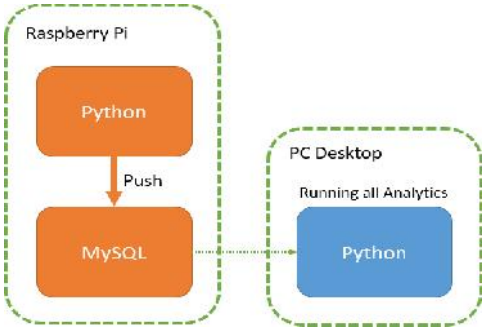
Total dataset yang didapatkan dari 4 variabel (*Timestamp*, Jarak, Kelembapan, dan Suhu) ini adalah 9981 dataset, yang diambil selama 1 minggu, tanpa ada pergerakan benda apapun.

B. *Raspberry Pi Specification*

Tipe Raspberry Pi yang digunakan adalah Raspberry Pi 3 dengan OS Debian 4.19.66-v7. Perangkat mini-PC ini memiliki Bahasa pemrograman Python dengan versi 3.5.3, dan MySQL dengan versi 3.4.

C. *Pengolahan Data*

Mini-PC ini akan melakukan running Python setiap 1 menit selama 1 minggu. Program Python sudah di setting untuk membaca dataset suhu, jarak, kelembapan, dan *timestamp* dan di push datanya ke MySQL di dalam Raspberry Pi. Pengolahan data ini akan kembali dihitung oleh Python di sisi PC Desktop yang telah disetting. Untuk membaca MySQL tersebut, Tools ini bisa memanfaatkan library Python yang sudah berbasis Data Sains, diantaranya adalah *Numpy*, *scikit*, *Pandas*, *Matplotlib*, dan *Seaborn*.



Gambar 2. Diagram komunikasi metode ini

D. Metode Pengolahan Data

Untuk mendeteksi *fraud* dalam penelitian ini, klasifikasi yang digunakan adalah *K-means* dan *Euclidean Distance*. Beberapa metode dalam penelitian yang dilakukan adalah:

1. *Data Preprocessing* dan *Data Cleansing* (deteksi data yang hilang, normalisasi data)
2. Menentukan optimum nilai *K* dalam metode *K-means*
3. Mengklasifikasi dataset menggunakan *K- means*
4. Melakukan perhitungan *Euclidean Distance* secara manual untuk mendeteksi *fraud* dari hasil simulasi klasifikasi *K-means*
5. Evaluasi hasil klasifikasi ini dengan menggunakan metode *Outlier*

III. HASIL PENELITIAN

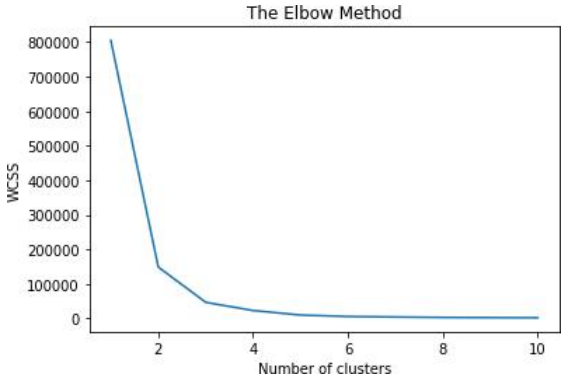
A. *Data Preprocessing*

Total Dataset yang didapatkan adalah 9981 dataset dengan 4 variabel (Jarak, Suhu, Kelembapan, *Timestamp*). Namun, seiring dengan performansi sensor dan *Hardware* raspberry pi, ada beberapa *missing* data sebanyak 16 dataset atau setara dengan ratio 0.1%, yang relatif tidak signifikan terhadap datasetnya. Oleh karenanya, *missing* informasi 16 dataset ini di isi dengan total rata-rata (*Mean*) dari dataset yang tersedia untuk menutupi nilai variabel *Not a Number (NaN)*.

B. *Elbow Method*

K-Means clustering adalah salah satu metode *unsupervised learning* dengan meletakkan sebuah *Centroid* secara random dan terpisah satu sama lain. Metode ini bertujuan untuk mengukur *Variance* antara *Centroid* dengan *dataset Observasi* yang berdekatan, atau biasa di sebut sebagai WCSS (*within-cluster sums of squares*) [5]. Jumlah *centroid* dalam klusterisasi *K-Means* ditentukan dari nilai K, dimana K adalah equivalen dengan total jumlah *Centroid*nya. Metode untuk menentukan Nilai K disebut sebagai *Elbow Method.*

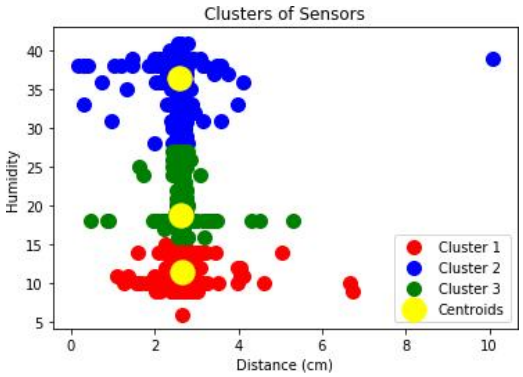
Nilai K terbaik adalah ketika nilai WCSS tersebut berada dalam titik awal pertama kali berubah menjadi nilai yang linear, dalam hal ini, sesuai dengan Gambar 2, nilai K terbaik adalah 3.



Gambar 3. Metode Elbow untuk menentukan banyaknya nilai K

C. *K-Means Clustering*

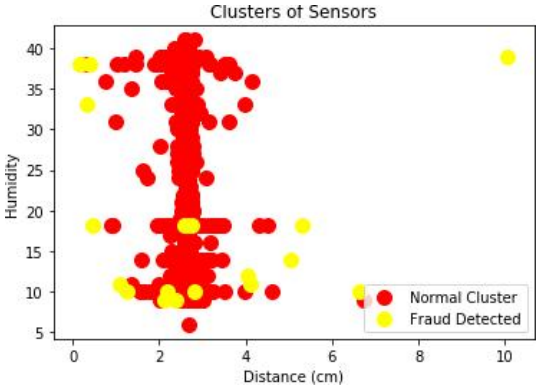
Nilai *K* ini akan diimplementasikan ke dalam sistem *K-means* library milik python, dengan nilai *K* = 3, dengan membandingkan dataset jarak dengan kelembapan. Dengan menggunakan 3 *centroid*, maka klasifikasi dataset ini akan dibagi menjadi 3 cluster, dengan algoritma *machine learning* yang digunakan. Hasil yang didapatkan untuk pembagian cluster dapat dilihat pada Gambar 3 ini.



Gambar 4. Pembagian cluster jarak vs kelembapan dengan metode K-Means

D. Euclidean Distance

Setelah menggunakan *K-Means Clustering* yang telah di bagi 3 cluster, peneliti mencoba mengoptimalisasi nilai pembagian cluster ini untuk mendapatkan deteksi kecurangan data terbaik. Gambar 4 menunjukkan pola persebaran Cluster Jarak vs kelembapan dengan menggunakan metode *Euclidean Distance + K-Means*.



Gambar 5. Pembagian cluster jarak vs kelembapan dengan metode *K-Means + Euclidean Distance*

IV. PEMBAHASAN

A. Perbandingan K-Means vs K-Means + Euclidean Distance

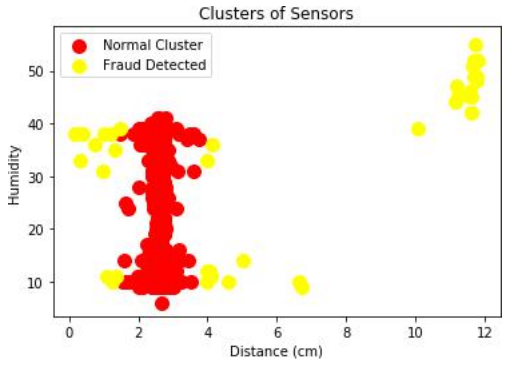
Gambar 3 dan Gambar 4 menunjukkan sebuah perbandingan pendeteksi *Fraud* antara metode *K- Means* dengan *Euclidean Distance*. Pada dasarnya, untuk metode *K-Means* ini hanya membagi dataset menjadi 3 cluster, tanpa mendeteksi *Fraud* tersebut. Dengan metode *Euclidean Distance*, hasil yang didapatkan relatif lebih baik untuk bisa mendeteksi Anomali data.

B. Evaluasi Hasil dengan Metode Outlier

Untuk melakukan evaluasi deteksi *Fraud* ini, peneliti mencoba untuk membandingkan metode *K- means + Euclidean Distance* dengan *outlier detection*. Lalu, evaluasi ini akan membandingkan perhitungan *confusion matrix*, dimana metode ini di support oleh *library* di Python, yang dapat membandingkan seberapa banyak kesamaan antara kedua metode ini.

Gambar 5 menunjukkan persebaran cluster antara *normal cluster* dan *Fraud Detected*, dengan menggunakan formula *outlier* :

- *Higher Layer : + 3 * Standard Deviation*
- *Lower Layer : - 3 * Standard Deviation*



Gambar 6. Pembagian cluster jarak vs kelembapan dengan metode *Outlier Detection*

Nilai hasil *Outlier Detection* dibandingkan kembali dengan tabel *Confusion Matix* sebagai metode untuk menentukan performansi model klasifikasi dari sebuah dataset yang ada. *Output* nilai dari *confusion matrix* ini dituangkan dalam 4 kategori :

1. *True Positive* (TP) : Ketika sebuah observasi bernilai positif dan prediksinya pun nilainya positif
2. *True Negative* (TN) : Ketika sebuah observasi bernilai negatif, dan prediksinya pun nilai negatif
3. *False Positive* (FP) : Ketika sebuah observasi bernilai negatif dan prediksinya berlawanan (positif)
4. *False Negative* (FN) : Ketika sebuah observasi bernilai positif dan prediksinya berlawanan (negatif)

Nilai komparasi *confusion matrix* antara kedua metode (*K-Means + Euclidean Distance*) dengan *Outlier Detection* adalah:

TABLE I
CONFUSION MATRIX TABLE

		Predicted	Predicted
		NO	YES
Actual	NO	TP = 9927	FP = 7
Actual	YES	FN = 17	TN = 17

Untuk menganalisis lebih jauh terkait evaluasi kedua metode ini, salah satu parameter yang dapat ditentukan dari hasil *confusion matriks* adalah:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} * 100\% = 99.7\% \quad (1)$$

Dari hasil evaluasi ini, nilai akurasi antara kedua metode ini adalah 99.7% dengan total dataset sebanyak 9965 (setelah adanya data cleansing). Hasil ini menunjukkan metode *K-means + Euclidean Distance* baik digunakan untuk mendeteksi *Fraud*.

V. PENUTUP

A. Kesimpulan

Sebuah *hardware* baik itu PC, *Router*, maupun *Microcontroller* yang dapat diakses secara publik memiliki kerentanan terhadap *Fraud* yang besar. Dalam penelitian ini, Raspberry Pi, salah satu mini- PC memiliki dataset yang variatif selama 1 minggu, menunjukkan sebuah anomali dalam dataset ketika di buka aksesnya secara publik via *Port Forwarding*.

Penelitian ini mencoba untuk mengklasifikasi nilai dataset yang teridentifikasi dengan *Fraud*, dengan metode *K-Means + Euclidean Distance*. Akurasi *K-Means + Euclidean Distance* ini adalah 99.7% dibandingkan dengan *outlier detection*, yang menunjukkan nilai yang relatif baik untuk menentukan deteksi anomali data.

B. Saran

Di penelitian berikutnya, penulis mencoba menggunakan sensor yang memiliki kinerja yang jauh lebih baik. Kinerja sensor sangat menentukan juga apakah data tersebut bisa bertahan untuk mengukur kalkulasi secara periodik. Selain itu, peneliti mencoba komparasi dengan metode lainnya, seperti *Artificial Neural Network* (ANN) dan *Naïve Bayes* untuk mengoptimalkan hasil ini.

DAFTAR PUSTAKA

[1] N. Malini and M. Pushpa, "Analysis on credit card fraud identification techniques based on KNN and outlier detection," *Proc. 3rd IEEE Int. Conf. Adv. Electr. Electron. Information, Commun. Bio- Informatics, AEEICB 2017*, pp. 255–258, 2017.

[2] A. Mishra and C. Ghorpade, "Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques," *2018 IEEE Int. Students' Conf. Electr. Electron. Comput. Sci. SCEECS 2018*, pp. 1–5, 2018.

[3] L. F. Gong, "The application of Naive Bayesian Classification in anti-fraud system of telecommunications," *Proc. - 2011 8th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2011*, vol. 2, pp. 1061–1064, 2011.
[4] K. Eremenko and H. de Ponteves, "Machine Learning A-Z™: Hands-On Python & R In Data Science," *SuperDataScience Team, Udemy*. [Online]. Available: https://www.udemy.com/machinelearning/.
[5] L. Minitab, "Interpret all statistics and graphs for Multiple Regression," 2019. [Online]. Available: https://support.minitab.com/en-us/minitab- express/1/help-and-how-to/modeling- statistics/regression/how-to/multiple- regression/interpret-the-results/all-statistics-and- graphs/#r-sq-adj.