ANALISA DAN PREDIKSI IKLAN LOWONGAN KERJA PALSU DENGAN METODE NATURAL LANGUAGE PROGRAMING DAN MACHINE LEARNING

Hary Sabita[1], Fitria[2], Riko Herwanto[3]

[123]Fakultas Ilmu Komputer, Informatics & Business Institute Darmajaya
Jl. 2.A. Pagar Alam No. 93, Bandar Lampung - Indonesia 35142
Telp. (0721) 787214 Fax. (0721) 700261
e-mail: hary.sabita@darmajaya.ac.id, fitria@darmajaya.ac.id, rikoherwanto@darmajaya.ac.id

ABSTRACT

This research was conducted using the data provided by Kaggle. This data contains features that describe job vacancies. This study used location-based data in the US, which covered 60% of all data. Job vacancies that are posted are categorized as real or fake. This research was conducted by following five stages, namely: defining the problem, collecting data, cleaning data (exploration and pre-processing) and modeling. The evaluation and validation models use Naïve Bayes as a baseline model and Small Group Discussion as end model. For the Naïve Bayes model, an accuracy value of 0.971 and an F1-score of 0.743 is obtained. While the Stochastic Gradient Descent obtained an accuracy value of 0.977 and an F1-score of 0.81. These final results indicate that SGD performs slightly better than Naïve Bayes.

Keywords—*NLP, Machine Learning, Naïve Bayes, SGD, Fake Jobs*

ABSTRAK

Selama masa pandemi Covid 19, penipuan lowongan pekerjaan atau ketenaga kerjaan meningkat sehingga tekanan ekonomi dan dampak virus Corona telah secara signifikan mengurangi ketersediaan pekerjaan dan kehilangan pekerjaan bagi banyak orang. Kasus seperti ini menghadirkan peluang yang tepat bagi penipu. Kebanyakan *scammer* melakukan ini untuk mendapatkan informasi pribadi dari orang yang mereka *scam*. Penelitian ini menggunakan sample data dari lokasi yang berbasis di AS, yang mencakup 60% dari seluruh data. Lowongan pekerjaan yang diposting, dikategorikan sebagai nyata atau palsu. Penelitian ini dilakukan dengan mengikuti lima tahap, yaitu: pendefinisian masalah, pengumpulan data, pembersihan data (eksplorasi dan pra-pemrosesan) serta pemodelan. Model evaluasi dan validasi menggunakan *Naïve Bayes* sebagai model *baseline* dan *Stochastic Gradient Descent* sebagai model akhir. Untuk model *Naïve Bayes* didapatkan nilai akurasi sebesar 0,971 dan skor F1 sebesar 0,743. Sementara pada *Stochastic Gradient Descent* didapatkan nilai akurasi sebesar 0,977 dan skor F1 sebesar 0,81. Hasil akhir ini menunjukkan bahwa *Stochastic Gradient Descent* memiliki kinerja yang sedikit lebih baik dibandingkan *Naïve Bayes*.

Kata Kunci — NLP, *Machines Learning, Naïve Bayes, Stochastic Gradient Descent*, Lowongan Pekerjaan Palsu

I. PENDAHULUAN

Selama masa pandemi Covid 19, penipuan lowongan pekerjaan atau ketenagakerjaan meningkat. Menurut *Consumer News and Business Channel* (CNBC), jumlah penipuan ketenagakerjaan meningkat dua kali lipat dibandingkan dengan 2017 [1]. Situasi pasar saat ini telah menyebabkan tingginya pengangguran. Tekanan ekonomi dan dampak virus Corona secara signifikan mengurangi ketersediaan pekerjaan dan kehilangan pekerjaan bagi banyak orang. Kasus seperti ini menghadirkan peluang yang tepat bagi penipu [2-3]. Kebanyakan *scammer* melakukan ini untuk mendapatkan informasi pribadi dari orang yang mereka *scam* [4]. Informasi pribadi dapat berisi alamat, rekening bank, nomor jaminan sosial, dan lain-lain [5]. Para penipu memberi pencari kerja peluang kerja yang sangat menguntungkan dan kemudian meminta uang sebagai imbalan. Atau mereka membutuhkan investasi dari pencari kerja dengan janji pekerjaan. Ini adalah masalah berbahaya yang dapat diatasi melalui teknik *Machine Learning* dan *Natural Language Processing* (NLP) [6-10].

II. METODE PENELITIAN

a. Algoritma dan Teknik

Berdasarkan analisis data yang diperoleh, pemodelan akhir akan menggunakan data teks maupun numerik. Sebelum melakukan pemodelan data, maka kumpulan data akhir ditentukan. Penelitian ini menggunakan kumpulan data dengan fitur-fitur analisis akhir sebagai berikut:

1. *Telecommuting*
2. Kecurangan (*Fraudulent*)
3. Rasio; rasio pekerjaan palsu dan nyata berdasarkan lokasi.
4. Teks; kombinasi dari judul, lokasi, profil_perusahaan, deskripsi, persyaratan, manfaat, pengalaman_wajib, pendidikan_wajib, industri serta fungsi.

Sementara untuk algoritma dan teknik yang digunakan pada penelitian ini adalah:

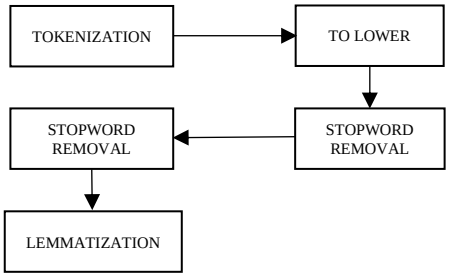
1. NLP
2. Algoritma *Naïve Bayes*
3. Pengklasifikasi *Stochastic Gradient Descent (*SGD*)*

Proses yang dilakukan adalah, *Naïve Bayes* dan pengklasifikasi SGD dibandingkan untuk melihat akurasi dan skor F1, selanjutnya model akhir ditentukan. Pada penelitian ini, *Naïve Bayes* digunakan untuk menghitung probabilitas kondisional dua peristiwa berdasarkan probabilitas kemunculan setiap peristiwa. Pengklasifikasian SGD

mengimplementasikan rutinitas pembelajaran penurunan gradien stokastik sederhana yang mendukung fungsi kerugian dan pinalti berbeda untuk klasifikasi. Model ini digunakan pada teks dan numerik secara terpisah, kemudian hasil akhir digabungkan.

b. Metodologi

Untuk pemrosesan data, langkah – langkah yang dilakukan adalah:



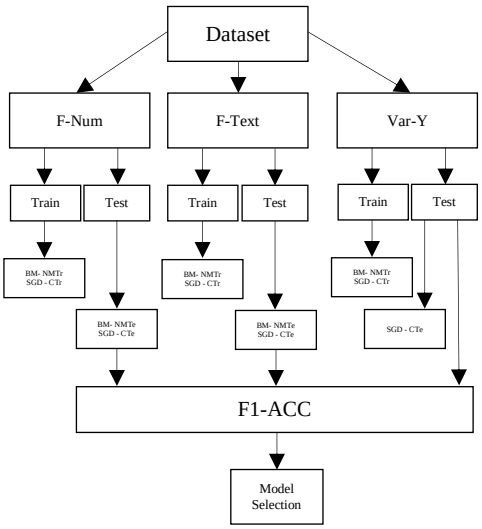
Gambar 1. II. Pemrosesan Data

- *Tokenization*: Data tekstual dibagi menjadi beberapa unit yang lebih kecil. Data dipecah – pecah menjadi kata – kata.
- *To Lower*: Kata – kata yang dipisahkan, diubah menjadi huruf kecil.
- Stopword Removal: Adalah kata – kata yang tidak menambahkan banyak arti pada kalimat, seperti: the, a, an, he, have dan lain – lain.
- *Lemmatization*: Proses pengelompokan lemmatisasi dimana bentuk kata – kata yang berubah digunakan bersama – sama.

c. Implementasi

Dataset dibagi menjadi teks, numerik dan variabel y. Dataset teks diubah menjadi matriks frekuensi *term* untuk analisis lebih lanjut. Dengan menggunakan *free software*, dan memungkinkan kita melakukan beragam pekerjaan dalam *Data Science*, seperti regresi, klasifikasi, pengelompokkan,

data *preprocessing*, *dimensionality reduction* dan *model selection* (pembandingan, validasi, dan pemilihan parameter maupun model) atau *Scikit learning*, dataset tersebut dipecah menjadi dataset *test* dan *training*. Model dasar *Naïve Bayes* dan model SGD dilatih menggunakan set *train*, sekitar 70% dari kumpulan data. Hasil akhir dari model, berdasarkan dua set pengujian – numerik dan teks, digabungkan sehingga jika kedua model mengatakan bahwa titik data tertentu bukan hanya penipuan, maka lowongan kerja tersebut adalah palsu. Hal ini dilakukan untuk mengurangi bias algoritma *Machine Learning* terhadap kelas mayoritas. Model yang dilatih digunakan pada set pengujian, dengan tujuan untuk mengevaluasi performa model. Akurasi dan skor F1 dari *Naïve Bayes* dan SGD dibandingkan, selanjutnya pemilihan model akhir untuk analisis.



Gambar 2. Proses Pemilihan Model Akhir

Keterangan:
F-Num = Fitur numerik (Telecommuting, Rasio, Jumlah Karakter)
F-Text = Fitur text (idf dan jumlah matriks
Var-Y = Variabel Y (Kecurangan-Fraudulent)
BM-NMTr = *Baseline Model – Naïve Model Training*
SGD-CTr = *Stochastic Gradient Descent Classifier Training*
BM-NMTe = *Baseline Model – Naïve Model Test*
SGD-CTe = *Stochastic Gradient Descent Classifier Test*
F1-ACC = F1 and Accuracy Calculation and Comparison

d. Metrik

Penelitian ini melakukan evaluasi dengan menggunakan dua metrik, yaitu:

1. Akurasi

Rumus metrik ini didefinisikan sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Keterangan:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Rumus ini menghasilkan rasio dari semua poin data yang dikategorikan dengan benar ke semua poin data. Hal ini sangat berguna untuk mengidentifikasi lowongan pekerjaan nyata atau palsu. Kelemahan pada metrik ini adalah, algoritma *Machine Learning* cenderung menyukai kelas yang dominan. Data yang ada pada penelitian ini kelasnya tidak seimbang, akurasi yang tinggi akan menunjukkan seberapa baik model kita kategorikan sebagai kategori negatif (lowongan pekerjaan nyata).

2. Skor F1

Ini adalah ukuran akurasi model pada kumpulan data. Dijelaskan pada rumus (2):

$$Skor\ F1 = \frac{TP}{TP+(\frac{FP+FN}{2})} \quad (2)$$

Keterangan:
TP = True Positive
FP = False Positive
FN = False Negative

Skor F1 digunakan, karena dalam skenario ini, negatif palsu dan positif palsu sangat penting. Model ini diperlukan untuk mengidentifikasi kedua kategori dengan skor setinggi mungkin, karena biaya keduanya sangat tinggi.

III. HASIL DAN PEMBAHASAN

a. Eksplorasi Data

Pada penelitian ini, dataset yang digunakan berjumlah 17.880 data dan 18 fitur. Tabel 1, menerangkan fitur – fitur dataset yang digunakan, serta *script* informasi untuk data fitur.

```
In [1]: import pandas as pd

In [6]: lowonganPalsu = pd.read_csv('fake_job_postings.csv')

In [7]: lowonganPalsu.columns

Out[7]: Index(['job_id', 'title', 'location', 'department', 'salary_range',
       'company_profile', 'description', 'requirements', 'benefits',
       'telecommuting', 'has_company_logo', 'has_questions', 'employment_type',
       'required_experience', 'required_education', 'industry', 'function',
       'fraudulent'],
      dtype='object')
```

Tabel 1. Fitur-Fitur Dataset yang Digunakan

No	Variabel	Tipe data	Keterangan
1	job_id	int	nomor identifikasi lowongan
2	title	text	nama jabatan atau pekerjaan
3	location	text	lokasi pekerjaan
4	department	text	departemen pekerjaan
5	salary_range	text	gaji yang ditawarkan
6	company_profile	text	informasi perusahaan
7	description	text	ringkasan perusahaan
8	requirements	text	informasi kebutuhan awal pekerjaan
9	benefits	text	keutungan yang ditawarkan
10	telecommuting	boolean	informasi pekerjaan remote atau on site
11	has_company_logo	boolean	apakah perusahaan memiliki logo
12	has_questions	boolean	apakah pekerjaan memiliki pertanyaan
13	employment type	text	kategori pekerjaan
14	required_experience	text	kebutuhan pengalaman pekerjaan
15	required_education	text	kebutuhan jenjang pendidikan
16	industry	text	jenis industri
17	function	text	fungsi pekerjaan secara umum
18	fraudulent	boolean	0; nyata, 1; palsu

Karena sebagian besar tipe data ini adalah teks, maka tidak digunakan statistik ringkasan. Tipe data *integer* hanya ada pada job_id, dan tidak relevan untuk analisis ini. Selanjutnya dataset dieksplorasi untuk mengidentifikasi nilai *null.* Variabel yang banyak memiliki nilai yang hilang adalah department dan salary_range. Untuk nilai salary_range yang hilang, diisi dengan nilai rata – rata dari kolom tersebut. Berikut *script* yang informasinya.

```
In [11]: lowonganPalsu.isnull().sum()
Out[11]: job_id                   0
         title                    0
         location               346
         department           11547
         salary_range         15012
         company_profile       3308
         description              1
         requirements          2695
         benefits              7210
         telecommuting            0
         has_company_logo         0
         has_questions            0
         employment_type       3471
         required_experience   7050
         required_education    8105
         industry              4903
         function              6455
         fraudulent               0
         dtype: int64
```

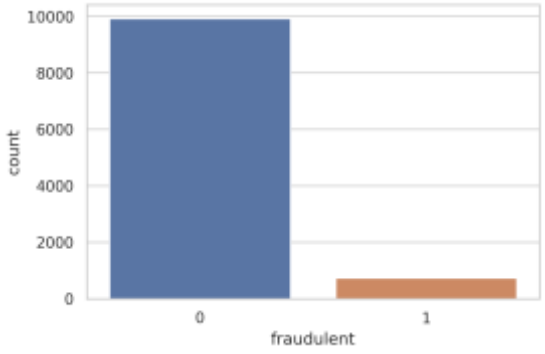
Dataset ini berisi lowongan pekerjaan dari beberapa negara, dalam rentan tahun 2020. Banyak data menggunakan bahasa yang berbeda, data yang digunakan adalah data yang berbasis lokasi di AS, yang mencakup 60% dari kumpulan data. Data ini dijadikan sampel karena semua dalam bahasa Inggris, agar mudah ditafsirkan. Selain itu, lokasi dibagi menjadi negara bagian dan kota untuk analisis lebih lanjut. Setelah dilakukan eksplorasi data, maka dataset akhir memiliki sampel sebanyak 10.656. Gambar 3. menjelaskan plot perhitungan untuk lowongan pekerjaan

yang menghasilkan 9926 (93,14%) lowongan pekerjaan nyata dan hanya 730 (6,85%) lowongan pekerjaan palsu, serta *script* untuk jumlah lowongan yang ada di AS.

```
In [41]: lowonganPalsu_US = lowonganPalsu[lowonganPalsu['location'].\
                                 str.contains('US')]

In [42]: len(lowonganPalsu_US)
Out[42]: 10656

In [43]: len(lowonganPalsu_US[lowonganPalsu_US.fraudulent == 0]),\
         len(lowonganPalsu_US[lowonganPalsu_US.fraudulent == 1])
Out[43]: (9926, 730)
```

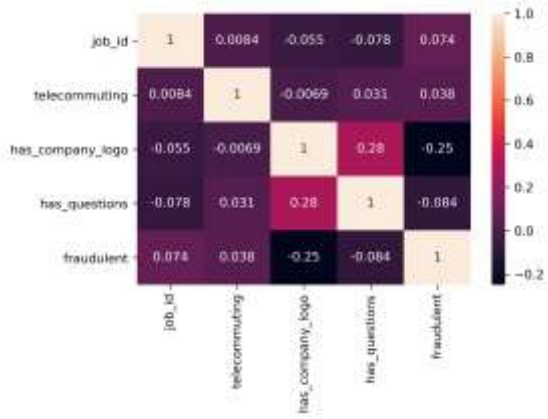


Gambar 3. Jumlah Lowongan Kerja Nyata dan Palsu Di AS

b. Analisis Eksplorasi Data

Langkah pertama yang dilakukan untuk memvisualisasikan dataset pada penelitian ini adalah dengan cara membuat matriks korelasi. Tujuannya adalah untuk mempelajari hubungan data numerik. Matriks korelasi tidak menunjukkan korelasi positif atau negatif yang kuat antara data numerik. Gambar 4. menjelaskan hubungan korelasi matriks pada dataset dan informasi *script* nya.

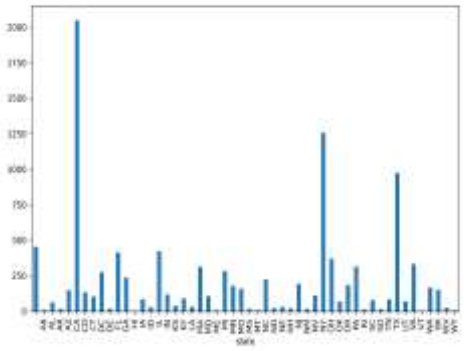
```
In [87]: sns.heatmap(lowonganPalsu_US.corr(), annot=True);
```



Gambar 4. Korelasi Hubungan Matriks pada Dataset

Selanjutnya fitur yang dieksplorasi adalah fitur tekstual. Eksplorasi data ini dimulai dari lokasi yang ada di negara bagian AS. Gambar 5., jumlah pekerjaan berdasarkan lokasi dan informasi *script* nya.

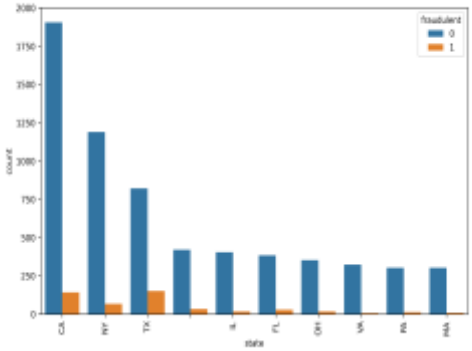
```
In [93]: plt.figure(figsize=(10,6))
         lowonganPalsu_US.groupby('state').fraudulent.count().plot(kind='bar');
         plt.savefig('lokerAS.png', dpi=300)
         plt.show()
```



Gambar 5. Jumlah Lowongan Pekerjaan Berdasarkan Lokasi

Grafik pada Gambar 5., menunjukkan bahwa negara bagian; California, New York dan Texas memiliki jumlah lowongan pekerjaan tertinggi. Untuk mengeksplorasi data lebih jauh, Gambar 6, menampilkan distribusi jumlah pekerjaan nyata dan palsu di 10 negara bagian teratas beserta informasi *script* nya.

```
In [92]: plt.figure(figsize=(10,6))
         sns.countplot(x='state', data=lowonganPalsu_US, hue="fraudulent", \
                       order=lowonganPalsu_US['state'].\
                       value_counts().iloc[:10].index)
         plt.xticks(rotation=90)
         plt.savefig('linear826.png', dpi=300)
         plt.show()
```



Gambar 6. Distribusi Lowongan Pekerjaan Nyata Dan Palsu Di 10 Negara Bagian

Hasilnya menunjukkan bahwa Texas dan California memiliki kemungkinan lowongan pekerjaan palsu lebih tinggi dibandingkan negara bagian lainnya.

c. Hasil Evaluasi dan Validasi Model

Model terakhir yang digunakan untuk analisis ini adalah SGD. Hal ini dilakukan berdasarkan hasil Pengukuran yang dibandingkan dengan model dasar. Nilai perbandingan akurasi dan skor F1 antara Naïve Bayes (model *baseline*) dan SGD (model akhir), disajikan pada Tabel 2, berikut dengan informasi *script* nya.

```
In [224]: from sklearn.naive_bayes import MultinomialNB
          from sklearn import metrics
          from sklearn.model_selection import train_test_split

          nb_classifier = MultinomialNB()
          nb_classifier.fit(count_train, y_train)
          pred = nb_classifier.predict(count_test)

In [225]: metrics.accuracy_score(y_test, pred) # Nilai akurasi untuk Naive Bayes/ model baseline
Out[225]: 0.9711950010366065

In [227]: metrics.f1_score(y_test, pred) # Nilai Skor F1 untuk Naive Bayes/ model baseline
Out[227]: 0.7435897435897436
```

```
In [222]: from sklearn.linear_model import SGDClassifier

          clf_log = SGDClassifier(loss='log').fit(count_train, y_train)
          pred_log = clf_log.predict(count_test)

In [223]: metrics.accuracy_score(y_test, pred_log)
Out[223]: 0.9776869787133547

In [226]: clf_num = SGDClassifier(loss='log').fit(X_train_num, y_train)
          pred_num = clf_num.predict(X_test_num)
          metrics.accuracy_score(y_test, pred_num)
Out[226]: 0.9336384439359268

In [227]: prediction_array = []

          for i, j in zip(pred_num, pred_log):
              if i == 1 and j == 0:
                  prediction_array.append(0)
              else:
                  prediction_array.append(j)

In [228]: metrics.accuracy_score(y_test, prediction_array) # Nilai Akurasi untuk SGD
Out[228]: 0.9768306493615604

In [230]: metrics.f1_score(y_test, prediction_array) # Nilai skor F1 untuk SGD
Out[230]: 0.8146453089244851
```

Tabel 2. Perbandingan Nilai Akurasi dan Skor F1

Model	Akurasi	Skor F1
Naïve Bayes	0,971	0,743
SGD	0,977	0,81

IV. SIMPULAN

Berdasarkan nilai akhir yang didapatkan, model SGD memiliki kinerja lebih baik dari model *Naïve Bayes*. Model SGD dapat mengidentifikasi pekerjaan nyata dengan akurasi yang sangat tinggi sebesar 0,977 dan Skor F1 sebesar 0,81.

PENELITIAN LANJUTAN

Dataset yang digunakan dalam penelitian ini sangat tidak seimbang. Kebanyakan lowongan pekerjaannya adalah nyata dan sedikit sekali yang palsu. Karena itu, pekerjaan nyata diidentifikasi dengan cukup baik. Untuk itu, penelitian berikutnya diharapkan dapat menambahkan teknik yang dapat menghasilkan sampel kelas minoritas,

sehingga dataset menjadi seimbang dan memberikan hasil yang lebih baik.

DAFTAR PUSTAKA

[1] Habiba, S. I. (2021). A Comparative Study on Fake Job Post Prediction Using Different Data Mining Technique. *In 2021 2nd International Conference on Robotics, Electrical and Signaling Processing Techniques (ICREST)*, 543-546.

[2] Ranparia, D., Kumari, S., & Sahani, A. (2020, November). Fake Job Prediction using Sequential Network. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)* (pp. 339-343). IEEE.

[3] Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet*, *9*(1), 6.

[4] Zamir, H. (2020). Cybersecurity and Social Media. *Cybersecurity for Information Professionals: Concepts and Applications*.

[5] Wilson, J. (2018). Scamming the scammers with their own tricks. *Computer Fraud & Security*, *2018*(9), 14-16.

[6] B Keerthana, A. R. (2021). *Accurate Prediction of Fake Job Offers Using Machine Learning.* Singapore: Springer.

[7] Ahmed, A. A. A., Aljabouh, A., Donepudi, P. K., & Choi, M. S. (2021). Detecting Fake News using Machine Learning: A Systematic Literature Review. *arXiv preprint arXiv:2102.04458*.

[8] Ranparia, D., Kumari, S., & Sahani, A. (2020, November). Fake Job Prediction using Sequential Network. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)* (pp. 339-343). IEEE.

[9] Mahbub, S. &. (2018). Using Contextual Features for Online Recruitment Fraud Detection. *Association For Information Systems*, -.

[10] Oshikawa, R., Qian, J., & Wang, W. Y. (2018). A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

[11] Sabita, H., & Herwanto, R. (2020). Pantauan Prediktif Covid-19 Dengan Menggunakan Metode SIR

dan Model Statistik Di Indonesia. *TEKNIKA*, *14*(2), 145-150.

[12] Wibowo, H., & Indriyani, F. (2018, October). K-Nearest Neighbor Method For Monitoring Of Production And Preservation Information (Treatment) Of Rubber Tree Plant. In *International Conference on Information Technology and Business (ICITB)* (pp. 29-44).

[13] Herwanto, R., Purbo, O. W., & Sriyanto, S. (2020). Membandingkan Performa Antara Hyperledger Dan Mysql. *Jurnal Informatika*, *20*(1), 89-100.