Jurnal Manajemen Informatika Jayakarta
Volume 3, Nomor 3, (Juli 2023): halaman 187-193

P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v3i3.1114

PEMBUATAN MODEL DATA MINING UNTUK MEMPREDIKSI INFEKSI PENYAKIT COVID-19 DENGAN MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER

[1]Yogie Wilvren Saragih, [2] Samuel Krispama Lumbantoruan, [3] Ali Muhammad Saleh Baaboud , [4] EvaYulia Puspaningrum

Program Studi Informatika, Fakultas Ilmu Komputer
Universitas Pembangunan Nasional "Veteran" Jawa Timur

Email : 20081010050@student.upnjatim.ac.id, 20081010066@student.upnjatim.ac.id, verbal598@gmail.com, 20081018261@student.upnjatim.ac.id,

ABSTRAK

Pandemi Covid-19 telah menjadi masalah kesehatan masyarakat yang kompleks dan membutuhkan pemrosesan cepat dan kerja sama solusi di berbagai bidang. Teknik preprocessing data digunakan untuk membersihkan data dan menghilangkan outlier dan noise. Algoritma Naive Bayes kemudian diterapkan untuk penambangan data untuk mengklasifikasikan kasus positif dan negatif dan mengelompokkan data berdasarkan fitur-fiturnya.Hasilnya menunjukkan keefektifan algoritma Naive Bayes dalam penambangan data Covid-19, dengan akurasi tinggi dalam mengklasifikasikan kasus positif dan negatif. Analisis pengelompokan juga mengungkapkan perbedaan signifikan kasus Covid-19 antar provinsi di Indonesia.Hasilnya dapat digunakan untuk mendukung pengambilan keputusan oleh pemerintah dan masyarakat dalam penanganan pandemi Covid-19 di Indonesia.

Kata kunci: *Covid-19, Naïve Bayes, Pandemi*

ABSTRACT

The Covid-19 pandemic has become a complex public health problem and requires fast processing and collaborative solutions in various fields. Data preprocessing techniques are used to clean data and remove outliers and noise. Naive Bayes algorithm is then applied for data mining to classify positive and negative cases and group data based on their features. The results show the effectiveness of the Naive Bayes algorithm in data mining for Covid-19, with high accuracy in classifying positive and negative cases. Clustering analysis also reveals significant differences in Covid-19 cases between provinces in Indonesia. The results can be used to support decision making by the government and society in handling the Covid-19 pandemic in Indonesia.

Keywords: *Covid-19, Naïve Bayes, Pandemic*

1 PENDAHULUAN

Pada era pandemi Covid-19, prediksi infeksi yang akurat menjadi sangat penting untuk membantu mengendalikan penyebaran virus. Oleh karena itu, pembuatan model data mining untuk memprediksi infeksi penyakit Covid-19 telah menjadi topik yang menarik dalam penelitian ilmiah. Salah satu

Jurnal Manajemen Informatika Jayakarta
Volume 3, Nomor 3, (Juli 2023): halaman 187-193
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v3i3.1114

algoritma yang digunakan dalam memprediksi infeksi penyakit Covid-19 adalah algoritma Naive Bayes Classifier.

2 TINJAUAN PUSTAKA

Pada penelitian mengenai Pembuatan model data mining untuk memprediksi infeksi penyakit Covid-19 dengan menggunakan Algoritma Naive Bayes Classifier, Jupyter Notebook digunakan untuk melakukan beberapa tahap pemrosesan data, seperti data cleaning, data preprocessing, dan feature selection. Selain itu, Jupyter Notebook juga digunakan untuk membangun dan mengevaluasi model Naive Bayes Classifier.

Dalam penelitian ini, Jupyter Notebook digunakan untuk melakukan beberapa tugas, antara lain:

- Data cleaning: Data Covid-19 yang didapatkan dari sumber-sumber terpercaya seringkali memiliki format yang tidak teratur dan mengandung data yang tidak relevan. Oleh karena itu, Jupyter Notebook digunakan untuk membersihkan data dengan menghapus data yang tidak relevan dan mengubah format data menjadi format yang sesuai.

- Data preprocessing: Data Covid-19 yang telah dibersihkan perlu diproses lebih lanjut sebelum dapat digunakan untuk membangun model Naive Bayes Classifier. Beberapa teknik preprocessing yang dapat dilakukan antara lain normalisasi data, penghapusan outlier, dan pengkodean variabel kategorikal. Jupyter Notebook digunakan untuk melakukan teknik preprocessing yang sesuai dengan karakteristik data.

- Feature selection: Sebelum membangun model Naive Bayes Classifier, perlu dilakukan pemilihan fitur atau feature selection untuk menentukan variabel mana yang paling relevan dalam memprediksi infeksi Covid-19. Jupyter Notebook digunakan untuk melakukan beberapa teknik feature selection seperti mutual information, chi-square, dan correlation-based feature selection.

- Membangun model Naive Bayes Classifier: Setelah melakukan preprocessing dan feature selection, langkah selanjutnya adalah membangun model Naive Bayes Classifier. Jupyter Notebook digunakan untuk mengimplementasikan algoritma Naive Bayes Classifier dan melatih model menggunakan data training.

- Evaluasi performa model: Setelah model dibangun, perlu dilakukan evaluasi performa model untuk menentukan seberapa akurat model dalam memprediksi infeksi Covid-19. Jupyter Notebook digunakan untuk melakukan evaluasi performa model menggunakan beberapa metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.

Dalam penelitian atau project data mining, penggunaan Jupyter Notebook sangat membantu dalam melakukan eksplorasi data, preprocessing, pemilihan fitur, dan pembangunan model. Dengan menggunakan Jupyter Notebook, pengguna dapat melihat hasil dari setiap tahap pemrosesan data secara interaktif dan dapat memperbaiki kode dengan cepat dan mudah. Selain itu, Jupyter Notebook juga memungkinkan pengguna untuk membuat dokumentasi yang rapi dan mudah dipahami.

Jurnal Manajemen Informatika Jayakarta
Volume 3, Nomor 3, (Juli 2023): halaman 187-193
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v3i3.1114

2.1 Metode Naïve Bayes

Algoritma Naive Bayes Classifier merupakan salah satu metode klasifikasi data mining yang berdasarkan pada teorema Bayes. Algoritma ini dapat digunakan untuk memprediksi kelas dari sebuah data berdasarkan pada kemungkinan kelas tersebut terjadi. Dalam konteks prediksi infeksi penyakit Covid-19, algoritma Naive Bayes Classifier dapat digunakan untuk memprediksi kemungkinan seseorang terinfeksi berdasarkan pada faktor-faktor seperti usia, jenis kelamin, riwayat perjalanan, dan gejala yang dialami.

3 HASIL DAN PENGUJIAN

Dengan menggunakan data yang didapat dari kaggle , maka proses analisa dilakukan sesuai tahapan berikut ini :

3.1 Mengkoneksikan dataset ke dalam Jupyter Notebook

```
import pandas as pd
import numpy as np
dataset =  pd.read_csv("covid19.csv")
dataset.head()
```

Dan tampilan dataframe, seperti tampak pada gambar berikut ini :

```
In [20]: dataset
```

Out[20]:

	Sesak Napas	Demam	Batuk Kering	Sakit Tenggorokan	Hidung Meler	Sakit Kepala	Kelelahan	Gastrointestinal	Kontak Dengan Pasien COVID	Menghadiri Pertemuan Besar	Keluarga Yang Bekerja di Tempat Umum	Terinfeksi-COVID
0	YES	YES	YES	YES	YES	NO	YES	YES	YES	NO	YES	YES
1	YES	YES	YES	YES	NO	YES	YES	NO	NO	YES	NO	YES
2	YES	YES	YES	YES	YES	YES	YES	YES	NO	NO	NO	YES
3	YES	YES	YES	NO	NO	NO	NO	NO	NO	YES	NO	YES
4	YES	YES	YES	YES	YES	YES	NO	YES	YES	NO	NO	YES
...	...	...	...	...	...	...	...	...	...	...	...	...
5429	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	NO	YES
5430	YES	YES	YES	NO	YES	YES	YES	NO	NO	NO	NO	YES
5431	YES	YES	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO
5432	YES	YES	YES	NO	YES	YES	NO	NO	NO	NO	NO	NO
5433	YES	YES	YES	NO	YES	YES	YES	NO	NO	NO	NO	NO

5434 rows × 12 columns

3.2 Melihat detail informasi mengenai tipe data yang ada di dataframe sebelum data diolah

```
In [26]: dataset.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5434 entries, 0 to 5433
Data columns (total 12 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Sesak Napas                           5434 non-null   int64
 1   Demam                                 5434 non-null   int64
 2   Batuk Kering                          5434 non-null   int64
 3   Sakit Tenggorokan                     5434 non-null   int64
 4   Hidung Meler                          5434 non-null   int64
 5   Sakit Kepala                          5434 non-null   int64
 6   Kelelahan                             5434 non-null   int64
 7   Gastrointestinal                      5434 non-null   int64
 8   Kontak Dengan Pasien COVID            5434 non-null   int64
 9   Menghadiri Pertemuan Besar            5434 non-null   int64
 10  Keluarga Yang Bekerja di Tempat Umum  5434 non-null   int64
 11  Terinfeksi-COVID                      5434 non-null   int64
dtypes: int64(12)
memory usage: 509.6 KB
```

Jurnal Manajemen Informatika Jayakarta
Volume 3, Nomor 3, (Juli 2023): halaman 187-193
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v3i3.1114

3.3 Mendeklarasikan vektor fitur dan variabel yakni Terinfeksi COVID

```
# Variabel independen
x = dataset.drop(["Terinfeksi-COVID"], axis = 1)
# Variabel dependen
y = dataset["Terinfeksi-COVID"]
```

```
In [28]: print(x)
```

```
      Sesak Napas  Demam  Batuk Kering  Sakit Tenggorokan  Hidung Meler  \
0               1      1             1                  1             1
1               1      1             1                  1             0
2               1      1             1                  1             1
3               1      1             1                  0             0
4               1      1             1                  1             1
...           ...    ...           ...                ...           ...
5429            1      1             0                  1             1
5430            1      1             1                  0             1
5431            1      1             1                  0             0
5432            1      1             1                  0             1
5433            1      1             1                  0             1

      Sakit Kepala  Kelelahan  Gastrointestinal  Kontak Dengan Pasien COVID  \
0                0          1                 1                           1
1                1          1                 0                           0
2                1          1                 1                           0
3                0          0                 0                           0
4                1          0                 1                           1
...            ...        ...               ...                         ...
5429             0          1                 1                           0
5430             1          1                 0                           0
5431             0          0                 0                           0
5432             1          0                 0                           0
5433             1          1                 0                           0

      Menghadiri Pertemuan Besar  Keluarga Yang Bekerja di Tempat Umum
0                              0                                     1
1                              1                                     0
2                              0                                     0
3                              1                                     0
4                              0                                     0
...                          ...                                   ...
5429                           0                                     0
5430                           0                                     0
5431                           0                                     0
5432                           0                                     0
5433                           0                                     0

[5434 rows x 11 columns]
```

3.4 Memisahkan dataset menjadi training dan testing set

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state=42)
x_train.shape, x_test.shape, y_train.shape, y_test.shape
```

Daftar Rujukan

3.5 Hasil encode variabel kategorikal (X_test) dan Hasil encode variabel kategorikal (X_train)

Jurnal Manajemen Informatika Jayakarta
Volume 3, Nomor 3, (Juli 2023): halaman 187-193
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v3i3.1114

```
In [31]: x_test
```

Out[31]:

	Sesak Napas	Demam	Batuk Kering	Sakit Tenggorokan	Hidung Meler	Sakit Kepala	Kelelahan	Gastrointestinal	Kontak Dengan Pasien COVID	Menghadiri Pertemuan Besar	Keluarga Yang Bekerja di Tempat Umum
3372	1	1	1	0	0	0	1	1	0	1	1
4950	0	1	0	1	1	0	0	1	0	0	0
2146	0	1	1	1	0	0	1	0	0	1	1
501	1	1	1	1	0	0	1	0	1	1	1
3418	1	1	1	0	1	1	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
4576	1	1	1	1	1	0	1	0	0	0	0
4871	0	1	0	1	1	1	0	1	0	0	0
4472	0	0	1	0	1	0	0	1	0	0	1
3814	1	0	1	0	0	0	1	1	0	1	0
4313	0	1	0	0	1	1	1	1	0	0	0

1359 rows × 11 columns

```
In [40]: x_train
```

Out[40]:

	Sesak Napas	Demam	Batuk Kering	Sakit Tenggorokan	Hidung Meler	Sakit Kepala	Kelelahan	Gastrointestinal	Kontak Dengan Pasien COVID	Menghadiri Pertemuan Besar	Keluarga Yang Bekerja di Tempat Umum
1263	1	1	1	1	1	1	0	0	0	1	1
5087	0	0	1	0	0	0	0	0	1	0	0
999	1	1	1	1	1	0	1	0	1	1	0
5146	1	0	1	1	0	0	1	0	1	1	0
3573	0	1	1	0	1	1	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...
3772	0	1	1	0	1	1	1	1	0	1	1
5191	1	1	0	1	1	1	0	0	1	1	0
5226	1	1	0	0	1	0	0	1	1	1	0
5350	1	0	0	0	0	1	0	0	1	1	0
860	1	1	1	1	0	0	0	0	1	1	0

4075 rows × 11 columns

3.6 Menghitung model Categorical NB(), menentukan x_train dan x_test

```
In [32]: from sklearn.naive_bayes import CategoricalNB
         classifier = CategoricalNB()
         classifier.fit(x_train, y_train)
         classifier.class_count_
Out[32]: array([ 792., 3283.])

In [33]: y_pred = classifier.predict(x_test)
         y_pred
Out[33]: array([1, 0, 1, ..., 0, 1, 0], dtype=int64)

In [34]: np.array(y_test)
Out[34]: array([1, 0, 1, ..., 0, 1, 0], dtype=int64)
```

Jurnal Manajemen Informatika Jayakarta
Volume 3, Nomor 3, (Juli 2023): halaman 187-193
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v3i3.1114

3.7 Menghitung matrix model

```
In [36]: from sklearn.metrics import confusion_matrix
         cm = confusion_matrix(y_test, y_pred)
         print(cm)

         [[ 216   43]
          [  15 1085]]
```

3.8 Menghitung nilai akurasi klasifikasi Naive Bayes

from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))

	precision	recall	f1-score	support
0	0.94	0.83	0.88	259
1	0.96	0.99	0.97	1100
accuracy			0.96	1359
macro avg	0.95	0.91	0.93	1359
weighted avg	0.96	0.96	0.96	1359

3.9 Prediksi gejala

```
In [41]: pasien_baru=np.array([[0,1,0,1,1,0,0,1,0,0,0]])
         prediksi_baru = classifier.predict(pasien_baru)

         if prediksi_baru == 0:
             print('Anda tidak memiliki gejala COVID-19. Stay At Home')
         else:
             print('Anda mungkin terinfeksi virus COVID-19! Silakan lakukan tes PCR')

         Anda tidak memiliki gejala COVID-19. Stay At Home
```

4. KESIMPULAN DAN SARAN

Dalam tugas akhir ini, model Naive Bayes Classifier memprediksi bahwa gejala kasus covid-19 didap at dari data 5434 dan model Gaussian Naive Bayes menunjukkan kinerja yang sangat baik dengan nila i akurasi model sebesar 95.73215599705665% serta nilai precision (0,96), recall (0,99) dan f-1 score (0,97).

DAFTAR PUSTAKA

[1] Pratiwi, R. W., & Nugroho, Y. S. 2016. Prediksi Rating Film Menggunakan Metode NaiveBayes. Jurnal Teknik Elektro, 8(2), pp. 60-63. doi: *https://doi.org/10.15294/jte.v8i2.7764*.

[2] Wasiati, H., & Wijayanti, D. Sistem Pendukung Keputusan Penentuan Kelayakan Calon Tenaga Kerja Indonesia Menggunakan Metode Naive Bayes (Studi Kasus: Di P.T. Karyatama MitraSejati Yogyakarta). IJNS – Indonesian Journal on Networking and Security, 3(2), pp. 45-51. doi: http://dx.doi.org/10.1123/ijns.v3i2.154.

P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v3i3.1114

[3] World Health Organization. (2023). Coronavirus . Retrieved from World Health Organization: *https://www.who.int/healthtopics/coronavirus*

[4] Fajar, Muhammad. 2020. Estimasi Angka Reproduksi Novel Coronavirus (COVID-19) Kasus Indonesia. Retrieved from https://www.researchgate.net/publication/340248900_ESTIMATION_OF_COVID_19_REPRODUCTIVE_NUMBER_CASE_OF_INDONESIA_Estimasi_Angka_Reproduksi_Novel_Coronavi rus_COVID-19_Kasus_Indonesia. doi: 10.13140/RG.2.2.32287.92328 (diakses tanggal 05 Juli 2023)

[5] Hikmah, N. (2017). Pemanfaatan Metode Naïve Bayes Classifier dalam Pembuatan SistemPakar untuk Diagnosa Penyakit Kelamin. EnergyJurnal Ilmiah Ilmu-Ilmu Teknik, 7(2), 50-55.

[6] Samsir, S., Ambiyar, A., Verawardina, U., Edi, F., & Watrianthos, R. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan MetodeNaïve Bayes. Jurnal Media Informatika Budidarma, 5(1), 157-163.