# Penerapan Algoritma Naïve Bayes Untuk Prediksi Penyakit Paru-Paru Menggunakan Rapid Miner

## Application Of The Naïve Bayes Algorithm For Prediction Of Lung Diseases Using Rapidminer

**[1]Yusuf Muhyidin, [2]Muhammad Rafi Muttaqin, [3]Imam Ma'ruf Nugroho, [4]Moch. Hafid**
[123]Teknik Informatika, STT Wastukancana
[1]yusufmuhyidin@wastukancana.ac.id, [2]rafi@wastukancana.ac.id, [3]imam.ma@wastukancana.ac.id,
[4]hafid@wastukancana.ac.id
Corresponding author: yusufmuhyidin@wastukancana.ac.id

Abstrak; Paru-paru adalah organ utama pada sistem pernapasan manusia yang terletak di dalam rongga dada dan berjumlah sepasang. Paru-paru merupakan organ vital yang sangat mempengaruhi kesehatan tubuh, karena memiliki fungsi untuk menjaga keseimbangan asam basa tubuh,mengelurakan karbon dioksida yang tidak dibutuhkan tubuh dan uap air. Merokok menjadi penyebab utama penyakit paru-paru, pada tahun 2020 berdasarkan laporan *World Health Organization* (WHO), diperkirakan 10 juta orang menderita penyakit paru-paru di seluruh dunia. Penelitian ini menggunakan klasifikasi algoritma naïve bayes untuk mendapatkan model prediksi yang dapat memprediksi data pasien penyakit paru-paru. Penelitian ini bertujuan untuk mendapatkan nilai akurasi dengan menggunakan algoritma naïve bayes. Adapun data yang digunakan pada penelitian ini diperoleh dari Kaggle yang berisi 469 data dengan 14 atribut di dalamnya. RapidMiner digunakan sebagai tools untuk menguji dataset pasien yang digunakan sehingga menghasilkan sebuah prediksi dengan tingkat akurasi sebesar 99,9% *risk false* (tidak berisiko memiliki penyakit paru-paru).

Kata Kunci: Paru-paru, Rapidminer, Naïve Bayes, Data Mining

*Abstract; The lungs are the main organ in the human respiratory system which is located in the chest cavity and consists of a pair. The lungs are a vital organ that greatly influences the body's health, because it has the function of maintaining the body's acid-base balance, removing carbon dioxide that the body does not need and water vapor. Smoking is the main cause of lung disease, in 2020 based on the World Health Organization (WHO) report, it is estimated that 10 million people suffer from lung disease worldwide. This research uses the Naïve Bayes classification algorithm to obtain a prediction model that can predict lung disease patient data. This research aims to obtain accuracy values using the Naïve Bayes algorithm. The data used in this research was obtained from Kaggle which contains 469 data with 14 attributes in it. RapidMiner is used as a tool to test the patient dataset used to produce a prediction with an accuracy rate of 99.9% risk false (no risk of having lung disease).*
*Keywords : Lung, Rapidminer, Naïve Bayes, Data Mining*.

## 1 Pendahuluan

Paru-paru adalah organ utama pada sistem pernapasan manusia yang terletak didalam rongga dada dan berjumlah sepasang. Paru-paru merupakan organ vital yang sangat mempengaruhi kesehatan tubuh, karenamemiliki fungsi untuk menjaga keseimbangan asam basa tubuh, mengelurakan karbondioksida yang tidak dibutuhkan tubuh dan uap air. Penyakit paru-paru adalah peradangan yang bisa mengganggu saluran pernapasan, seperti sesak napas, batuk, atau nyeri dada.Ada beberapa jenis penyakit paru-paru yangsering terjadi seperti pneumonia, Tuberkulosis (TBC), Bronkitis, penyakit paru obstruktif kronis, dan asma. Penyakit paru-paru adalah suatu penyakit yang dapat diobati dan disembuhkan. Namun membutuhkan waktu pengobatan dengan jangka waktu yang cukup lama. Penyebab risiko penyakit paru-paru selain dari asap tembakau, yaitu

polusi udara, bahan kimia dan debu pekerjaan, dan infeksi saluran pernapasan bawah yang sering terjadi selama massa kanak-kanak.

Pada tahun 2020 berdasarkan laporan *World Health Organization* (WHO), diperkirakan 10 juta orang menderita penyakit paru-paru di seluruh dunia. 5,6 juta pria dan 3,3 juta wanita, dan 1,1 juta anak-anak. Jumlah kasus terbanyak penyakit paru paru, yaitu 43% ad di Kawasan Asia Tenggara, diikuti Kawasan Afrika dengan 25%, dan 18% di Kawasan Pasifik Barat. Lebih dari 95% kasus dan kematian terjadi di negara-negara berkembang. Delapan negara menyumbangkan dua pertiga kasus penyakit paru-paru baru yaitu India, Tiongkok, Indonesia, Filipina, Pakistan, Nigeria, Bangladesh, dan Afrika Selatan.

Di seluruh dunia, hampir satu dari dua rumah tangga terdampak penyakit paru-paru menanggung biaya sebesar lebih dari 20% pemasukan rumah tangga. Meski perkembangangan di dunia medis begitu 49 pesat, namun masih banyak masyarakat yang tidak bisa menikmati perkembangannya. Contohnya hanya satu dari tiga orang yang mengidap Tuberkulosis Resistan Obat (TB-RO) yang mengakses pengobatan. Pendanaan di negara-negara berpendapatan rendah dan menengah yang menyumbangkan 98% kasus penyakit paru-paru dilaporkan masih jauh lebih rendah dari pendanaan yang dibutuhkan untuk pencegahan, diagnosis, pengobatan, dan pelayanan.

Berdasarkan penjelasan yang sudah diuraikan di atas, peneliti menggunakan Teknik data mining dengan algoritma Naïve Bayes dan metode *Knowledge Discovery in Database* (KDD) untuk memprediksi penyakit paru-paru. Dataset diiperoleh dari Kaggle yang berisi 469 data dengan 14 atribut di dalamnya.

## 2 Kajian Pustaka

### 2.1 Data mining

Data Mining adalah teknik pengolahan suatu pengetahuan yang didasarkan pada sebuah big data,data yang biasa digunakan diambil dari *database, warehouse data*, web dan lain-lain untuk diproses menjadi informasi yang menarik. Data miningsendiri banyak digunakan di beberapa bidang untuk membantu peningkatan dalam hal pengetahuan, selain itu juga data mining digunakan untuk meningkatkan penjualan. (Febriani dan Sulistiani, 2021). Cabang ilmu lain mendukung data mining yang mencakup statistik, teknologi basis data, *machine learning*, sistem pakar, algoritma paralel, algoritma genetika, pengenalan pola, visualisasi data, dan lain sebagainya (Anam dan Santoso, 2018).

Data Mining merupakan teknologi yang sangat berguna untuk membantu perusahaan menemukan informasi yang sangat penting dari gudang data mereka yang selama ini tidak diketahui apa manfaatnya.(Karlina dkk, 2022)

### 2.2 Paru-Paru

Paru-paru merupakan organ respirasi yang sangat penting bagi keberlangsungan hidup manusia. Paru-paru berfungsi menukar oksigen dari udara dengan karbondioksida dari darah. Jika kesehatan paru-paru terganggu maka fungsi kesehatan tubuh juga akan ikut terpengaruh secara keseluruhan (Musa dan Alang, 2017).

Paru-paru merupakan salah satu organ pada sistem pernapasanyang berfungsi sebagai tempat bertukarnya oksigen dengan karbondioksida di dalam darah. Gangguan paru-paru ini menyebabkanpenderita sulit bernafas, sulit beraktivitas, kekurangan oksigen bahkan apabila tidak cepat terdeteksi dapat menyebabkan kematian (Reni Rahma dewi & Rahmadi Kurnia, 2016). Menurut Departemen Kesehatan Republik Indonesia, penyakit paru termasuk salah satu penyakit yang kritis hingga saat ini.

### 2.3 Naïve Bayes

Klasifikasi adalah proses persiapan pencarian model yang menggambarkan data dan mengklasifikasikannya ke kelas-kelas yang ada.Naive Bayes adalah salah satu yang biasa digunakan karena keunggulan kecepatan dan akurasinya, naive bayes dianggap berpotensi sebagai metode klasifikasi data terbaik dalam hal akurasi dan perhitungan.

Naïve Bayes adalah suatu pengklasifikasian probabilitik sederhana yang menghitung peluang dari frekuensi dan kombinasi nilai dari dataset. Metode ini didasarkan pada asumsi bahwa nilai variable saling bebas jika diberikan nilai output. Naïve Bayes Classifier adalah suatu algoritma di dalam data mining yang menerapkan teorema Bayes untuk klasifikasi (Wulandari dkk, 2020). Persamaan teorema Bayes dituliskan sebagai berikut:

$$P(Y|X) = \frac{P(X|Y).P(Y)}{P(X)} \quad (1)$$

dimana:

P(Y|X) : Peluang terjadinya Y berdasarkan kondisi X (posteriori prob)
P(Y) : Peluang terjadinya Y (prior prob)
P(X|Y) : Peluang terjadinya X berdasarkan kondisi pada hipotesis Y
P(X) : Peluang terjadinya X

Karena nilai P(X) selalu tetap untuk setiap kelas pada suatu sampel yaitu bernilai 1, maka P(X) dapat dihilangkan. Sehingga, klasifikasi Naïve Bayes dapat dituliskan sebagai berikut:

$$\begin{aligned} P(Y|X_1, \dots, X_n) &= P(Y).P(X_1|Y).P(X_2|Y) \dots P(X_n|Y) \\ &= P(Y)\prod\prod_{i=1}^{n} P(X_i|Y) \end{aligned} \quad (2)$$

### 2.4 Rapid Miner

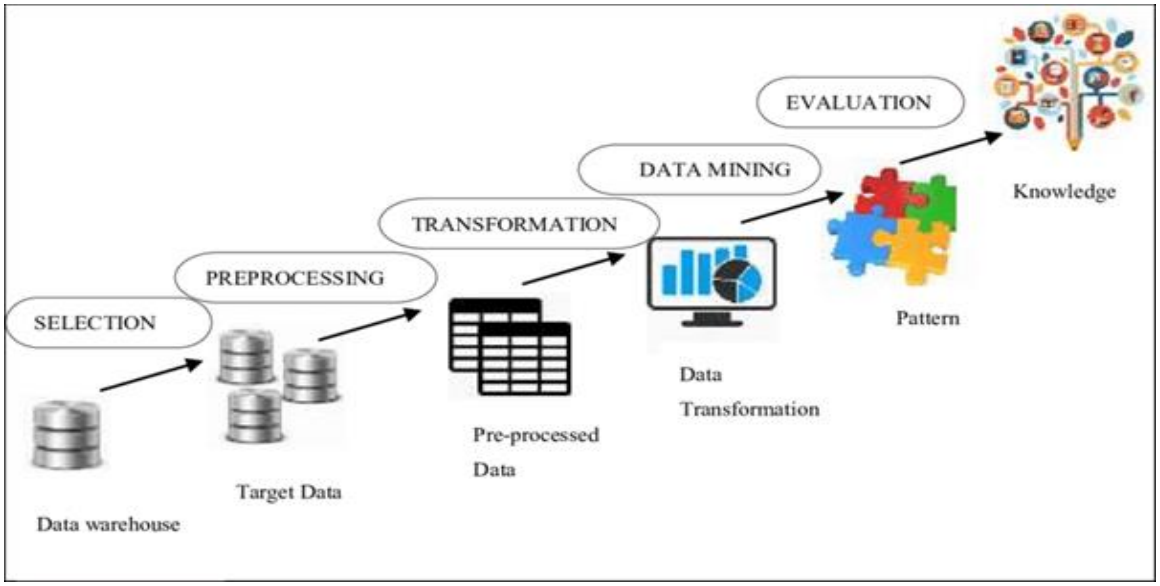
Rapid miner adalah sebuah aplikasi atau *software* yang berfungsi untuk mempelajari tentang data mining untuk analisis data, pemrosesan data, dan pengembangan model prediktif dan bersifat *open source*. Beberapa operator dalam rapid miner digunakan untuk membaca lembar kerja dari Microsoft excel. Setiap baris dalam tabel excel mewakili entitas data, sedangkan setiap kolom mewakili atribut data tersebut. Tampilan Rapid Miner yang ramah Pengguna memudahkan pengguna untuk menggunakanya Ketika rapidminer dijalankan, rapidminer akan menampilkan welcome prespective. Desgin prespective adalah tampilan kerja dari rapidminer, dan result prespective akan menampilkan hasil analisis. (Padilah dan Adam, 2019).

## 3 Metode *Knowledge Discovery in database* (KDD)

Metodologi pada penelitian ini menggunakan proses KDD (*knowledge discovery in database*). Proses ini menjelaskan secara sistematis dalam mencari suatu hubungan baru di dalam market basket analysis menggunakan beberapa tahap pengolahan data. Fayyad dkk. (1996) di dalam penelitiannya menjelaskan bahwa terdapat beberapa langkah di dalam proses KDD (*knowledge discovery in database*) diantaranya secara berurutan *selection, preprocessing, transformation, data mining, dan Interpretation/evaluation.*

a) ***Selection***: Data yang didapatkan dilakukan proses pemilihan terlebih dahulu. Dengan adanya data selection, proses pengolahan akan menjadi lebih baik sesuai dengan tujuan penelitian yang akan dicapai. Penelitian ini menggunakan 469 data dengan 14 atribut sumber data penyakit paru-paru didapatkan dari web kaggle

b) ***Preprocessing***: Merupakan beberapa proses persiapan data sebelum dilakukannya proses data mining. Data yang dilakukan *preprocessing* biasanya dilakukan dengan beberapa tahap seperti cleaning, reduction, integration.
c) ***Transformation***: Data harus dilakukan transformasi sebelum dilakukan pengolahan menggunakan data mining. Hal ini bertujuan untuk menyesuaikan data yang diolah berdasarkan algoritma dan *software* yang digunakan di dalam pengolahan data.
d) ***Data Mining***: Proses pengolahan data berdasarkan algoritma sesuai dengan teknik data mining. Algoritma yang digunakan pada penelitian yaitu Algoritma Naïve Bayes yang merupakan salah satu algoritma di dalam metode klasifikasi. Sehingga hasil luaran dari teknik ini berupa prediksi dari data yang diolah. Untuk proses pengolahan digunakan software RapidMiner Studio.
e) ***Interpretation/evaluation***: Merupakan proses menginterpretasikan hasil prediksi yang didapatkan dari teknik data mining. Pada bagian ini dilakukan percobaan hasil prediksi berdasarkan data training sebelumnya menggunakan data testing atau data yang belum diketahui kelasnya atau prediksinya.




**Gambar 1** Tahapan KDD
Sumber: (Turban, 2018)

## 4 Hasil dan Pembahasan

### 4.1 Data Penyakit Paru-Paru

Dataset yang diperoleh berjumlah 469 data, terdiri dari 70 data berisiko memiliki penyakit paru-paru dan 399 tidak berisiko memiliki penyakit paru-paru. Untuk lebih jelasnya dapat dilihat pada Gambar 2 dibawah ini:

**Tabel 1. Dataset Penyakit Paru-Paru**

| Patient | smoke | FVC | FEC1 | PEFR | O2 | ABG-P-O2 | ABG-P-CO2 | ABG-pH Leve | Scan | Asthama | therdisea | AGE | Risk |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Patient-1 | T | 2,85 | 2,16 | F | F | F | T | T | X-ray | F | F | 60 | F |
| Patient-2 | F | 3,4 | 1,88 | F | F | F | F | F | MRI | T | F | 51 | F |
| Patient-3 | F | 2,76 | 2,08 | F | F | F | T | F | X-ray | F | F | 59 | F |
| Patient-4 | F | 3.68 | 3,04 | F | F | F | F | F | X-ray | F | F | 54 | F |
| Patient-5 | F | 2,44 | 0,96 | F | T | F | T | T | X-ray | F | F | 73 | T |
| Patient-6 | F | 2,48 | 1,88 | F | F | F | T | F | X-ray | F | F | 51 | F |
| Patient-7 | F | 4,36 | 3,28 | F | F | F | T | F | MRI | T | F | 59 | T |
| Patient-8 | F | 3,19 | 2,5 | F | F | F | T | F | X-ray | F | F | 66 | T |
| Patient-9 | T | 3,16 | 2,64 | F | F | F | T | T | X-ray | F | F | 68 | F |
| Patient-10 | F | 2,32 | 2,16 | F | F | F | T | F | X-ray | T | F | 54 | F |

### 4.2 Data *Preprocessing*

Pada Preprocessing data merupakan tahapan yang dilakukan untuk mempersiapkan data sebelum nantinya akan digunakan dalam pemodelan. Langkah pertama yang dilakukan adalah pembersihan data dimana dilakukan pengecekan data hilang dan data duplikat. Hasil yang didapatkan menunjukkan bahwa tidak terdapat data hilang dan tidak terdapat data duplikat pada dataset yang digunakan.

### 4.3 Klasifikasi Naïve Bayes

Selanjutnya dilakukan proses klasifikasi menggunakan hitung manual, pertama mencari probabilitas priotnya sebagai berikut:

$P(C|X) = P(X|C)/P(C)$

P(Berisiko) = 70/469 = 0,149
P (Tidak berisiko) = 399/469 = 0,851

Sesudah dilakukan perhitung jumlah total kasus dari dataset penyakit paru-paru diperoleh nilai probabilitas prior untuk Berisiko adalah 1,149, sedangkan probalitas Tidak berisiko adalah 0,851.

Setelah melakukan Analisa pada data prediksi penyakit paru-paru dari dataset yang ada dengan menggunakan metode Naïve Bayes, maka hasil yang dicapai oleh peneliti adalah untuk mengetahui prediksi penyakit paru-paru seseorang berdasarkan gaya hidup, riwayat kesehatan dan atribut lainnya. Untuk mempermudah proses klasifikasi Naïve Bayes, selain menggunakan perhitungan

manual, peneliti juga menggunakan coding Naïve Bayes untuk mempercepat perhitungan dan sebagai alat komparasi hasil klasifikasi. Berdasarkan dataset yang terkumpul dilakukan proses perhitungan manual klasifikasi Naïve Bayes. Untuk lebih jelasnya dapat digunakan data uji dibawah ini:

Menghitung kelas Risk dengan perhitungan manual seperti berikut ini. Risk False = 399/469 = 0,851 MRI & Risk False = 217/399 = 0,544, Umur 51 & Risk False = 10/399 = 0,025, Other disease & Risk False = 3/399 = 0,008, Asthma & Risk False = 29/399 = 0,073, Smoke & Risk False = 48/399 = 0,120, FVC 4,72 & Risk False = 29/399 = 0,073, FEC1 3,56 & Risk False = 61/399 = 0,153, PEFR & Risk False = 24/399 = 0,060, O2 & Risk False = 54/399 = 0,135, ABG-P-O2 & Risk False = 23/399 = 0,058, ABG-P-CO2 & Risk False = 268/399 = 0,672, ABG-pH Level & Risk False = 61/399 = 0,153.

Semua atribut dikalikan maka hasil risk false adalah
0,851*0,544*0,025*0,008*0,073*0,120*0,073 *0,153*0,060*0,135*0,058*0,672*0,153 = **4,08E-13**

Hasil risk true adalah
0,149*0,557*0,029*0*0,143*0,129*0,029*0,0 86*0,100*0,214*0,129*0,786*0,243 = **0**

Hasil dari perkalian pada kelas Risk True didapatkan angka 0, sedangkan untuk kelas Risk False didapatkan angka 4,80E-13. Selanjutnya adalah mengkomparasi hasil dari Risk *True* dan Risk *False*. Disini hasil terbesar berada pada kelas *Risk False*, maka bisa diprediksi bahwa hasil data uji ini menghasilkan jawaban *Risk False* (Tidak berisiko memiliki penyakit paru-paru).

Setelah melakukan perhitungan manual, selanjutnya melakukan pemodelan menggunakan rapidminer dengan algpritma Naïve Bayes hasil dari pemodelan pada Gambar 2 dibawah ini:

ExampleSet (1 example, 3 special attributes, 14 regular attributes) Filter (1 / 1 examples): all

| Row No. | prediction(... | confidence(F) | confidence(T) | Patient | smoke | FVC | FEC1 | PEFR | O2 | ABG-P-O2 | ABG-P-CO2 | ABG-pH Lev.. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | F | 0.999 | 0.001 | Patient-470 | F | 4.720 | 3.560 | F | F | F | F | F |

**Gambar 2.** Hasil Model Prediksi Rapid Miner

Hasil pengolahan prediksi penyakit paru-paru mengunakan Aplikasi Naïve Bayes menghasilkan prediksi False atau Tidak berisiko memiliki penyakit paru-paru.

**SimpleDistribution**

```
Distribution model for label attribute Risk


Class F (0.851)
13 distributions

Class T (0.149)
13 distributions
```

Gambar 3. Hasil Probabilitas Prior

Pada Gambar 3, diperoleh kelas *Risk False* adalah 0,851, sedangkan kelas *Risk True* adalah 0,149. Dapat disimpulkan bahwa probabilitas prior pada perhitungan rapidminer dan manual sudah

sesuai. Dapat dilihat pada Gambar 3, tingkat akurasi pada dataset ini sangat tinggi yaitu sebesar 0,999 atau 99,9% *Risk False*. Maka dapat disimpulkan bahwa prediksi dari rapidminer untuk dataset ini memiliki akurasi yang baik.

## 5 Kesimpulan dan Saran

Paru-paru merupakan salah satu organ yang memiliki peran sebagai tempat bertukarnya oksigen dan karbon dioksida, namun ada kalanya paru-paru dapat mengalami kerusakan akibat gaya hidup tidak sehat. Merokok menjadi penyebab utama penyakit paru-paru, *World Health Organization* mengatakan lebih dari 40% kematian akibat merokok dan setidaknya 8 juta orang terbunuh setiap tahunnya. Penelitian ini mengambil data melalui situs penyedia dataset yaitu Kaggle sebanyak 469 data dengan 14 atribut Penelitian ini mengimplementasikan Teknik data mining klasifikasi yang merupakan salah satu teknik di dalam data mining. Algortima yang digunakan pada teknik ini adalah Algoritma Naive Bayes. Pada penerapannya, proses KDD (*knowledge discovery in database*) digunakan pada penelitian ini. Hasil prediksi algoritma Naïve Bayes dengan tingkat akurasi sebesar 99,9% risk false (tidak berisiko memiliki penyakit paru-paru). Aplikasi RapidMiner digunakan sebagai tools untuk menguji dataset pasien yang digunakan atau dataset testing.

Algoritma Naive Bayes yang digunakan pada penelitian ini menghasilkan akurasi yang baik untuk memprediksi penyakit paru-paru, untuk penelitian selanjutnya dapat dilakukan dengan mencoba algoritma lain dari teknik data mining agar mengetahui hasil akurasi yang di peroleh melalui algoritma lain sebagai perbandingan dengan penelitian ini.

**Referensi**

Chang Hartono dan Dwiyoga Widiantoro (2023) "Analisis Prediksi Harga Saham Unilever Menggunakan Regresi Linier dengan RapidMiner", Journal of Computer and Information Systems Ampera Vol. 5, No. 3, September 2023 e-ISSN: 2775-2496

C. Anam and H. B. Santoso (2018) “Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa,” Energy -J. Ilm. Ilmu-Ilmu Tek., vol. 8, no. 1, pp. 13–19.

Haris. (2022). Metode Naive Bayes UntukMemprediksi Penyakit Stroke. JurnalSistem Komputer dan Kecerdasan Buatan.

Karlena dkk (2022), "Market Basket Analysisuntuk Mengetahui Pola Beli KonsumenRoofboxMobilmenggunakan Algoritme Apriori", Jurnal Teknologika Volume 12 No. 2 November 2022, 308-316

Kaggle. (2024, April 25). Retrieved from https://kaggle.com

Musa dan Alang (2017) " Analisis Penyakit Paru-Paru Menggunakan Algoritmak-Nearest Neighbors Pada Rumah Sakit Aloei Saboekota Gorontalo" ILKOM Jurnal Ilmiah Volume 9 Nomor 3 Desember 2017, ISSN print 2087-1716, ISSN online 2548-7779

Reni Rahmadewi dan Rahmadi Kurnia (), "Klasifikasi Penyakit Paru Berdasarkan Citra Rontgen Dengan Metoda Segmentasi Sobel", ol: 5, No. 1, Maret 2016 ISSN: 2302 -294

S. Febriani and H. Sulistiani (2021) “Analisis Data Hasil Diagnosa Untuk Klasifikasi Gangguan Kepribadian Menggunakan Algoritma C4.5,” 89Jurnal Teknol. dan Sist. Inf., vol. 2, no.4, pp. 89–95, 2021

Turban, E., Delen, D., & Sharda, R. (2018). Business intelligence, analytics, and data science: A managerial perspective. Harlow; Munich: Pearson Prentice Hall.

T. N. Padilah and R. I. Adam, (2019) “Analisis Regresi Linier Berganda Dalam Estimasi Produktivitas Tanaman Padi Di Kabupaten Karawang,” FIBONACCI J. Pendidik. Mat. dan Mat., vol. 5, no. 2, p. 117, 2019, doi: 10.24853/fbc.5.2.117-128.

Wulandari, F., Jusia, P. A., & Jasmir. (2020). Klasifikasi Data Mining Untuk Mendiagnosa Penyakit ISPA Menggunakan Metode Naïve Bayes Pada Puskesmas Jambi Selatan. Jurnal Manajemen Teknologi Dan Sistem Informasi (JMS). 2(3): 214–227.