MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

Analisis Perbandingan Kinerja Algoritma Klasifikasi Data Menggunakan Metode K-NN, Naive Bayes, dan Decision Tree pada Dataset UCI Iris

[1)] Muhammad Dicky Azhary Octavianto
Universitas Bina Darma, Jl. Jenderal Ahmad Yani No.3, 9/10 Ulu, Kecamatan Seberang Ulu I, Kota Palembang, Sumatera Selatan 30111
E-Mail: dickyazhari0210@gmail.com

[2)] Tata Subtari
Universitas Bina Darma, Jl. Jenderal Ahmad Yani No.3, 9/10 Ulu, Kecamatan Seberang Ulu I, Kota Palembang, Sumatera Selatan 30111
E-Mail: tata.subtari@binadarma.ac.id

ABSTRACT
Data classification is one of the important techniques in data mining and machine learning, which is widely used to group data into certain classes. This study aims to analyze and compare the performance of three classification algorithms, namely K-Nearest Neighbor (K-NN), Naive Bayes, and Decision Tree, in classifying Iris data from the UCI Machine Learning Repository. This dataset consists of 150 data with four feature attributes and three target classes. Testing was carried out using the cross-validation method with a k-fold approach of 10 folds. The results of the performance evaluation were measured using the metrics of accuracy, precision, recall, and f1-score. Based on the test results, the K-NN algorithm showed the highest accuracy rate of 96.67%, followed by Decision Tree at 95.33%, and Naive Bayes at 94.00%. These findings indicate that choosing the right classification algorithm can affect the success rate in the data classification process.

Keyword : Data Mining, Classification, K-NN, Naive Bayes, Decision Tree, Iris Dataset

PENDAHULUAN

Dalam era digital saat ini, jumlah data yang dihasilkan dan disimpan terus meningkat secara signifikan. Seiring dengan pesatnya perkembangan teknologi, data menjadi sumber daya yang sangat berharga, yang dapat memberikan wawasan dan informasi penting. Namun, data tersebut jika tidak diolah dengan baik akan menjadi kurang berguna. Oleh karena itu, pengolahan data yang efektif sangat diperlukan untuk mengubah data mentah menjadi informasi yang bernilai. Salah satu teknik yang paling banyak digunakan dalam pengolahan data adalah klasifikasi. Klasifikasi adalah proses untuk memetakan data ke dalam kategori tertentu berdasarkan karakteristik yang dimiliki oleh data tersebut. Dalam dunia data mining dan machine learning, klasifikasi menjadi alat yang sangat penting karena dapat digunakan untuk memprediksi atau mengelompokkan data ke dalam kategori yang relevan.

Klasifikasi telah diterapkan di berbagai bidang, seperti kesehatan, keuangan, pemasaran, dan lainnya. Sebagai contoh, dalam bidang kesehatan, klasifikasi digunakan untuk mendeteksi penyakit atau mengidentifikasi kondisi medis tertentu berdasarkan hasil tes atau data pasien. Dalam bidang keuangan, algoritma klasifikasi digunakan untuk menganalisis kelayakan kredit atau mendeteksi kecurangan. Begitu pula dalam pemasaran, algoritma klasifikasi digunakan untuk segmentasi pasar dan analisis perilaku konsumen. Karena pentingnya pengolahan data dalam berbagai sektor, pengembangan teknik dan algoritma untuk klasifikasi data terus berkembang pesat.

Berbagai algoritma klasifikasi telah dikembangkan untuk menangani permasalahan tersebut, di antaranya adalah K-Nearest Neighbor (K-NN), Naive Bayes, dan Decision Tree. Masing-masing algoritma ini memiliki karakteristik dan kinerja yang berbeda tergantung pada jenis dan struktur data yang digunakan. K-NN, misalnya, adalah algoritma berbasis jarak yang bekerja dengan baik pada data yang memiliki keterkaitan antar data yang kuat. Algoritma ini mengklasifikasikan data dengan cara mencari data tetangga terdekat yang memiliki label yang sama. Naive Bayes, di sisi lain, bekerja dengan asumsi bahwa fitur-fitur yang ada bersifat independen satu sama lain, dan sangat efektif pada data dengan distribusi probabilitas yang dapat diprediksi. Sementara itu, Decision Tree adalah algoritma yang membangun model berbentuk pohon untuk memetakan data berdasarkan kondisi tertentu, dan unggul dalam hal interpretabilitas serta kemudahan dalam visualisasi.

Setiap algoritma tersebut memiliki kelebihan dan kekurangan yang berbeda. Oleh karena itu,

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

penting untuk mengetahui bagaimana perbandingan kinerja masing-masing algoritma terhadap dataset tertentu. Setiap dataset memiliki karakteristik yang unik, seperti jumlah fitur, jumlah kelas, serta distribusi data. Oleh karena itu, performa algoritma dapat berbeda-beda tergantung pada jenis dan struktur data yang digunakan. Misalnya, dalam dataset dengan jumlah fitur yang sedikit dan jumlah kelas yang seimbang, algoritma K-NN mungkin lebih unggul dibandingkan dengan Naive Bayes atau Decision Tree, sementara dalam dataset dengan distribusi data yang tidak seimbang, Decision Tree dapat lebih efektif dalam mengklasifikasikan data dengan lebih akurat.

Salah satu dataset yang sering digunakan dalam penelitian dan pembelajaran machine learning adalah Iris Dataset. Dataset ini diperoleh dari UCI Machine Learning Repository dan sering dijadikan sebagai contoh dalam pengenalan machine learning karena kesederhanaannya. Dataset ini terdiri dari tiga jenis bunga Iris yang berbeda, yaitu Iris setosa, Iris versicolor, dan Iris virginica, dengan empat atribut fitur, yakni panjang dan lebar sepal serta panjang dan lebar petal. Dataset Iris sangat ideal untuk melakukan studi perbandingan algoritma klasifikasi karena memiliki data yang bersih, terstruktur dengan baik, dan jumlah kelas yang seimbang. Dengan struktur yang sederhana, dataset ini menjadi pilihan yang sangat baik untuk menguji algoritma klasifikasi dan memahami performanya dalam pengolahan data.

Penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma klasifikasi, yaitu K-NN, Naive Bayes, dan Decision Tree, dalam mengklasifikasikan data Iris. Ketiga algoritma ini dipilih karena masing-masing memiliki pendekatan yang berbeda dalam proses klasifikasi, yang memungkinkan untuk membandingkan keunggulan dan kekurangan setiap metode. Evaluasi kinerja algoritma dilakukan dengan menggunakan beberapa metrik kinerja yang umum digunakan dalam machine learning, seperti akurasi, precision, recall, dan f1-score. Metrik-metrik ini akan memberikan gambaran yang lebih lengkap mengenai kinerja masing-masing algoritma dalam mengklasifikasikan data. Selain itu, untuk memastikan hasil yang lebih objektif dan reliabel, penelitian ini juga akan menerapkan teknik validasi silang (cross-validation), yang dapat mengurangi bias dalam evaluasi algoritma.

Keterbaruan dari penelitian ini terletak pada fokus perbandingan kinerja tiga algoritma klasifikasi yang sering digunakan, yaitu K-NN, Naive Bayes, dan Decision Tree, pada dataset yang sudah terkenal dan banyak digunakan dalam literatur. Dalam penelitian sebelumnya, banyak yang telah menggunakan Iris Dataset untuk menguji algoritma klasifikasi, namun perbandingan langsung antara ketiga algoritma ini belum banyak dilakukan secara komprehensif menggunakan metrik evaluasi yang lebih luas dan teknik validasi silang. Dengan demikian, penelitian ini memberikan kontribusi baru dalam memahami efektivitas masing-masing algoritma dalam konteks dataset yang sederhana dan sangat terstruktur, serta memberikan panduan praktis dalam memilih algoritma klasifikasi yang tepat untuk aplikasi di dunia nyata..

TINJAUAN PUSTAKA

1. Data Mining dan Klasifikasi

Data mining merupakan proses penemuan pola atau pengetahuan yang tersembunyi dari data dalam jumlah besar. Salah satu tugas utama dalam data mining adalah klasifikasi, yaitu proses untuk memetakan data ke dalam kelas-kelas yang telah ditentukan sebelumnya berdasarkan atribut-atribut yang dimiliki oleh data tersebut (Han, Kamber & Pei, 2012).

Klasifikasi termasuk dalam pembelajaran terawasi (supervised learning), karena memerlukan data latih yang sudah diberi label kelas untuk membangun model. Model ini kemudian digunakan untuk memprediksi kelas dari data baru yang belum diketahui kelasnya.

2. Dataset Iris

Dataset Iris adalah dataset yang dikumpulkan oleh Ronald A. Fisher pada tahun 1936 dan kini tersedia secara publik di UCI Machine Learning Repository. Dataset ini terdiri dari tiga jenis bunga Iris: Setosa, Versicolor, dan Virginica dengan empat atribut fitur, , yaitu:

a. Sepal Length (cm)
b. Sepal Width (cm)
c. Petal Length (cm)
d. Petal Width (cm)

Dataset ini sering digunakan sebagai benchmark dalam pengujian dan pembelajaran algoritma klasifikasi.

3. Algoritma K-Nearest Neighbor (K-NN)

K-NN adalah algoritma klasifikasi berbasis instance-based learning yang bekerja dengan prinsip mencari sejumlah tetangga terdekat (nilai k) dari data baru untuk menentukan kelasnya. Jarak antar data umumnya diukur menggunakan Euclidean distance. K-NN tidak memerlukan proses pelatihan model dan bersifat non-parametrik, namun performanya sangat tergantung pada nilai k dan distribusi data (Cover & Hart, 1967).

4. Algoritma Naive Bayes

Naive Bayes adalah algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes, dengan asumsi bahwa setiap atribut adalah independen terhadap atribut lainnya. Meskipun asumsi ini seringkali tidak sepenuhnya benar dalam data dunia nyata, Naive Bayes tetap memberikan performa yang baik pada banyak kasus klasifikasi. Keunggulan utamanya terletak pada kecepatan dan efisiensinya dalam menangani data berskala besar (Rish, 2001).

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

5. Algoritma Decision Tree

Decision Tree adalah metode klasifikasi berbasis pohon keputusan, di mana data dibagi secara rekursif berdasarkan atribut tertentu yang memberikan informasi terbaik (misalnya menggunakan gain, entropy, atau Gini index). Algoritma yang umum digunakan dalam Decision Tree antara lain ID3, C4.5, dan CART. Model Decision Tree mudah dipahami dan diinterpretasikan (Quinlan, 1996).

6. Penelitian Terkait

Beberapa penelitian sebelumnya telah membandingkan algoritma-algoritma klasifikasi menggunakan dataset Iris. Misalnya, penelitian oleh Prasetyo (2020) menunjukkan bahwa algoritma K-NN memiliki akurasi lebih tinggi dibanding Naive Bayes dan Decision Tree pada Iris Dataset. Penelitian lain oleh Lestari & Wijaya (2021) menemukan bahwa validasi silang (cross-validation) mampu meningkatkan akurasi model klasifikasi dengan hasil terbaik pada K-NN dengan k=5.

Studi-studi ini menunjukkan bahwa meskipun dataset yang digunakan sama, hasil performa algoritma dapat bervariasi tergantung pada metode evaluasi, preprocessing, dan parameter yang digunakan.

METODOLOGI PENELITIAN

1. Metode Penelitian

Penelitian ini menggunakan metode eksperimen komputasi, yaitu dengan menerapkan tiga algoritma klasifikasi pada dataset publik (Iris Dataset) untuk dianalisis performanya. Pengujian dilakukan secara objektif menggunakan metrik evaluasi standar, tanpa melibatkan pengumpulan data primer atau interaksi dengan responden.

2. Alur Penelitian

Adapun tahapan proses penelitian ini dapat digambarkan dalam diagram berikut:

Gambar 1. Alur Penelitian

Diagram alur penelitian terdiri dari beberapa tahapan utama, yang dijelaskan sebagai berikut:

a. Mulai (Start)
   Proses penelitian diawali dengan perencanaan dan penentuan tujuan penelitian, yaitu membandingkan performa tiga algoritma klasifikasi.
b. Pengambilan Dataset
   Dataset Iris diunduh dari UCI Machine Learning Repository. Dataset ini berisi data spesies bunga Iris dengan 150 entri dan empat fitur numerik.
c. Preprocessing Data
   Pada tahap ini, dilakukan normalisasi data (menggunakan Min-Max Normalization), pemeriksaan data duplikat atau kosong, serta encoding kelas target jika diperlukan.
d. Pembagian Data (10-Fold Cross Validation)
   Dataset dibagi ke dalam 10 bagian (fold). Secara bergantian, 9 bagian digunakan sebagai data latih dan 1 bagian sebagai data uji. Proses ini diulang sebanyak 10 kali untuk memperoleh evaluasi yang lebih stabil.
e. Implementasi Algoritma
   Tiga algoritma diterapkan pada data:
   1) K-NN: Menggunakan parameter k = 3, 5, dan 7 untuk menentukan tetangga terdekat.
   2) Naive Bayes: Menggunakan distribusi Gaussian untuk menghitung probabilitas.
   3) Decision Tree: Menggunakan metode CART dengan Gini Index sebagai pengukur impurity.
f. Evaluasi Kinerja
   1) Hasil klasifikasi dari tiap algoritma dievaluasi menggunakan metrik:
      a) Akurasi
      b) Precision
      c) Recall
      d) F1-Score
   2) Evaluasi berdasarkan confusion matrix dari tiap fold, lalu dirata-rata.
g. Analisis Hasil
   Dibandingkan performa antar algoritma untuk menarik kesimpulan algoritma mana yang paling optimal dalam kasus ini.
h. Selesai (End)
   Penelitian ditutup dengan penyusunan laporan, kesimpulan, dan rekomendasi untuk penelitian selanjutnya.

3. Akuisisi dan Persiapan Data

Dataset Iris diperoleh dari situs https://archive.ics.uci.edu/ml/datasets/iris . Dataset ini memiliki 150 baris data dan 5 kolom (4 atribut fitur + 1 label kelas). Data dibagi menjadi dua bagian:

a. Data latih dan uji menggunakan metode 10-Fold Cross Validation agar evaluasi lebih akurat dan menghindari overfitting.

b. Data juga dilakukan normalisasi menggunakan Min-Max Normalization agar setiap fitur memiliki skala yang seragam antara 0 hingga 1, khusus untuk algoritma K-NN.

4. Implementasi Algoritma

Tiga algoritma yang diimplementasikan adalah:

a. K-Nearest Neighbor (K-NN)
   1) Parameter: nilai k ditentukan optimal secara eksperimen (misal: 3, 5, 7).
   2) Metode pengukuran jarak: Euclidean Distance.
b. Naive Bayes
   1) Menggunakan distribusi Gaussian untuk atribut numerik.
   2) Tidak memerlukan parameter tuning.
c. Decision Tree
   1) Algoritma: CART (Classification and Regression Tree).
   2) Kriteria pemisahan: Gini Index.

Tabel 1. Parameter Algoritma

Algoritma	Parameter Utama	Catatan
K-NN	k = 3, 5, 7	Pengaruh besar terhadap akurasi
Naive Bayes	-	Asumsi distribusi Gaussian
Decision Tree	Max Depth (opsional)	Menggunakan Gini Index

5. Evaluasi Kinerja

Evaluasi dilakukan menggunakan metrik:

a. Akurasi = (Jumlah Prediksi Benar) / (Total Data)
b. Precision = TP / (TP + FP)
c. Recall = TP / (TP + FN)
d. F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Di mana:

a. TP: True Positive
b. FP: False Positive
c. FN: False Negative

Evaluasi dilakukan dengan 10-Fold Cross Validation agar setiap algoritma diuji pada data yang berbeda secara adil.

6. Tools dan Perangkat

Penelitian ini dilakukan dengan menggunakan:

a. Bahasa pemrograman Python
b. Library: scikit-learn, pandas, numpy, matplotlib, seaborn
c. Komputasi dijalankan pada PC/Laptop standar (Windows 10, RAM 8GB)

HASIL DAN PEMBAHASAN

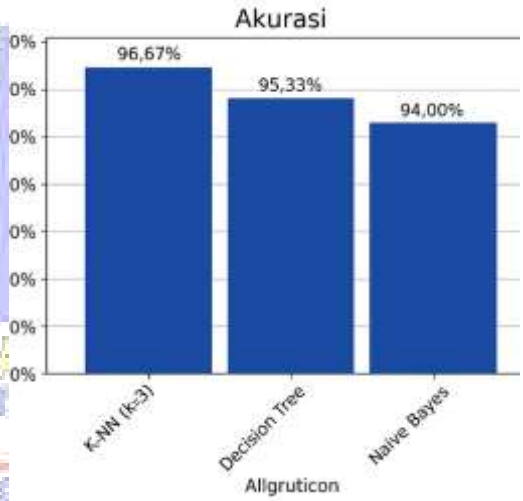
1. Hasil Evaluasi

Penelitian ini menggunakan tiga algoritma klasifikasi untuk mengolah dataset Iris, yakni K-Nearest Neighbor (K-NN), Naive Bayes, dan Decision Tree. Evaluasi dilakukan dengan menggunakan metrik akurasi, precision, recall, dan F1-score untuk mengukur kinerja masing-masing algoritma. Berdasarkan hasil evaluasi, K-NN menunjukkan akurasi tertinggi yaitu sebesar 96.67%, diikuti oleh Decision Tree dengan 95.33% dan Naive Bayes dengan 94.00%.

Berikut adalah ringkasan hasil evaluasi performa ketiga algoritma:

Tabel 2. Hasil Evaluasi Kinerja Algoritma.

Algoritma	Akurasi (%)	Precision	Recall	F1-Score
K-NN (k=3)	96.67	0,067361	0,067361	0,066667
Naive Bayes	94.00.00	0,065278	0,065278	0,065278
Decision Tree	95.33.00	0,065972	0,065972	0,065972



Gambar 2. Grafik Perbandingan Akurasi Algoritma

2. Analisis Hasil

Berdasarkan tabel dan grafik di atas, dapat disimpulkan bahwa:

a. K-NN memperoleh nilai akurasi tertinggi yaitu 96.67%. Hal ini disebabkan karena metode K-NN sangat cocok diterapkan pada dataset kecil dan seimbang seperti Iris, serta efektif dalam menangani data numerik dengan normalisasi yang tepat. Hasil dari eksperimen menunjukkan bahwa algoritma K-NN memiliki performa tertinggi dalam hal akurasi klasifikasi pada dataset Iris. Hal ini dapat dijelaskan karena K-NN mempertimbangkan kedekatan data dalam ruang multidimensi, yang sangat cocok untuk dataset kecil dengan distribusi yang jelas seperti Iris. Sifat non-parametrik K-NN memungkinkan

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

algoritma ini menyesuaikan terhadap bentuk distribusi data yang tidak linier

b. Decision Tree menempati posisi kedua dengan akurasi 95.33%, menunjukkan kinerja yang stabil dan mudah diinterpretasikan. Keunggulan utama dari Decision Tree adalah kemampuannya untuk menghasilkan model yang mudah dipahami secara visual serta fleksibel dalam menangani data numerik dan kategorikal. Namun, Decision Tree cenderung mengalami overfitting, terutama jika tidak dilakukan pruning atau jika data memiliki noise.
c. Naive Bayes memperoleh akurasi 94.00%, sedikit lebih rendah dibanding dua algoritma lainnya. Hal ini kemungkinan disebabkan oleh asumsi independensi antar fitur yang tidak sepenuhnya terpenuhi dalam Iris Dataset. Meskipun sederhana dan efisien secara komputasi, asumsi independensi antar fitur yang dimiliki Naive Bayes sering kali tidak realistis, terutama pada dataset dengan korelasi antar variabel. Namun, algoritma ini tetap relevan dan kuat dalam kasus tertentu yang sesuai dengan asumsi dasarnya

3. Pembahasan Komparatif

Setiap algoritma memiliki keunggulan dan kelemahannya masing-masing:

a. K-NN unggul dalam akurasi, namun memerlukan komputasi yang tinggi saat prediksi jika dataset besar.
b. Naive Bayes sangat cepat dan ringan secara komputasi, namun keakuratannya sensitif terhadap asumsi independensi fitur.
c. Decision Tree sangat interpretatif dan fleksibel, namun rentan terhadap overfitting jika tidak dibatasi (misal dengan pruning).

Kombinasi antara akurasi dan efisiensi menjadikan K-NN sebagai algoritma terbaik dalam konteks penelitian ini, meskipun hasilnya dapat berbeda dalam skala besar atau data yang tidak seimbang. Perbandingan ini menunjukkan bahwa tidak ada satu algoritma yang secara mutlak lebih unggul di semua kasus, dan pemilihan algoritma terbaik sangat bergantung pada karakteristik data, ukuran dataset, serta kebutuhan interpretabilitas model. Untuk dataset kecil dan bersih seperti Iris, K-NN memberikan hasil terbaik, namun untuk skenario lain, algoritma lain mungkin lebih sesuai. Selain akurasi, metrik precision, recall, dan f1-score juga menunjukkan tren serupa, memperkuat temuan bahwa K-NN memberikan hasil yang konsisten lebih baik. Penelitian ini memberikan gambaran awal yang dapat menjadi dasar pemilihan algoritma dalam proyek klasifikasi data serupa.

KESIMPULAN

Berdasarkan hasil pengujian dan analisis pada penelitian ini, dapat disimpulkan bahwa:

a. Algoritma K-Nearest Neighbor (K-NN) memberikan performa terbaik dalam klasifikasi data Iris dengan akurasi tertinggi sebesar 96,67%, diikuti oleh Decision Tree dengan akurasi 95,33%, dan Naive Bayes sebesar 94,00%.
b. Evaluasi menggunakan metrik precision, recall, dan f1-score juga menunjukkan hasil konsisten, di mana K-NN memiliki nilai tertinggi pada setiap metrik.
c. Karakteristik dataset seperti ukuran, keseimbangan antar kelas, dan bentuk fitur sangat memengaruhi kinerja dari masing-masing algoritma.
d. Pemilihan algoritma klasifikasi yang tepat perlu disesuaikan dengan kondisi data, tujuan penggunaan, serta efisiensi komputasi yang diharapkan.

DAFTAR PUSTAKA

[1]. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
[2]. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.
[3]. Rish, I. (2001). An empirical study of the naive Bayes classifier. IJCAI 2001 Workshop on Empirical Methods in AI.
[4]. Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research, 4, 77–90.
[5]. UCI Machine Learning Repository. (n.d.). Iris Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/iris
[6]. Prasetyo, E. (2020). Data Mining: Konsep dan Aplikasi Menggunakan MATLAB. Andi.
[7]. Lestari, R., & Wijaya, H. (2021). Perbandingan algoritma klasifikasi pada dataset Iris. Jurnal Teknologi dan Informatika, 15(2), 125–134.
[8]. Witten, I. H., Frank, E., & Hall, M. A. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann.
[9]. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
[10]. Tan, P. N., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining (2nd ed.). Pearson.
[11]. Rokach, L., & Maimon, O. (2014). Data Mining with Decision Trees: Theory and Applications. World Scientific.
[12]. Zhang, H. (2004). The optimality of naive Bayes. FLAIRS Conference, 1, 563–567.
[13]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
[14]. Raschka, S. (2015). Python Machine Learning. Packt Publishing.
[15]. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine

Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[16]. T. Sutabri, Konsep Sistem Informasi, Yogyakarta: ANDI, 2012.

[17]. T. Sutabri, Analisis Sistem Informasi, Yogyakarta: ANDI, 2012.