# IDENTIFIKASI ANOMALI DATA AKADEMIK MENGGUNAKAN DBSCAN *OUTLIER DETECTION*

**Estannisa Asfarina Fadlilah[1], Yulison Herry Chrisnanto[2], dan Ade Kania Ningsih[3]**
[1]Jurusan Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
Jl. Terusan Jend. Sudirman, Cibeber, Kec. Cimahi Selatan, Kota Cimahi, 40531
*Email : estandilli@gmail.com

**Abstrak**

*DBSCAN (Density Based Spatial Klastering of Aplikasi dengan Noise) adalah salah satu algoritma pengelompokan berbasis kepadatan. Pada penelitian ini memakai data akademik. Dengan mencari kelas atau klaster ideal pada metode DBSCAN terdapat banyak cara dalam menentukan hal tersebut. Salah satunya dengan metode Elbow. Hasil dari ini akan dijadikan dasar penentuan jumlah klaster dalam melakukan proses clustering dengan metode DBSCAN. Adanya outlier pada dataset sering dianggap sebagai salah perhitungan, oulier dapat membawa informasi yang signifikan atau informasi penting tidak ada pada data pengamatan yang dikumpulkan. Lebih parah lagi, outlier ini dapat berpengaruh pada pengambilan kesimpulan penelitian ingin di dapat berupa anomali. Clustering bisa digunakan sebagai metode deteksi outlier. Studi kasus ini menggunakan dataset mata kuliah inti jurusan informatika dari semester 1 sampai semester 6 pada angkatan 2014 – 2017 dengan nilai yang diambil hasil ulangan akhir semester (UAS) dan nilai akhir mata kuliah inti informatika di salah satu perguruan tinggi. Yang dimana terjadinya hasil deteksi anomali ini bisa mengetahui keberhasilan dari setiap mahasiswa karena dengan adanya deteksi ini bisa melihat hasil klaster mengarah pada titik yang baik atau mengarah pada titik yang kurang baik. Serta dapat mengevaluasi hasil dan memperbaiki pelajaran untuk keragaman penilaian mahasiswa yang berbeda.*

***Kata Kunci**: Algortima DBSCAN, Clustering, Data Mining, Deteksi Anomali, dan Outlier.*

## 1. PENDAHULUAN

Menurut (Novalia and Goejantoro, 2020), data *mining* adalah salah satu istilah yang digunakan untuk mengetahui pengetahuan yang tersembunyi di dalam database. Data *mining* merupakan proses semi otomatis yang menggunakan teknik statistik, matematika dan *mearching learning* untuk mengesktraksi dan mengidentifikasi informasi pengetahuan yang bermanfaat. Data *mining* adalah proses pencarian pola – pola yang tersembunyi berupa pegetahuan yang tidak diketahui sebelumnya dari sekumpulan data.

Pada algoritma DBSCAN menemukan nilai epsilon dan minpts untuk masing - masing klaster. Klaster yang memiliki banyak titik data akan dianggap sebagai klaster normal sedangkan klaster yang memiliki sedikit titik data akan dianggap sebagai klaster anomali. Pada fase deteksi, kami menetapkan titik baru ke klaster dan membuat peringatan jika poin ditetapkan ke klaster anomali. Juga memperbaharui perilaku normal dengan membuat klaster baru atau memperbaharui ukuran klaster.

Pada penelitian sebelumnya (Mayadewi and Rosely, 2015) salah satu indikator keberhasilan mahasiswa D3 Manajemen Informatika terhadap mata kuliah mereka. Mata kuliah tersebut adalah algoritma dan pemrograman, perancangan basis data, analisis dan perancangan sistem serta pemrograman web. Dimana tujuan utama dari studi ini adalah menggunakan data mining untuk memprediksi nilai proyek akhir mahasiswa berdasarkan nilai-nilai matakuliah yang mendukung proyek akhir mereka. Kemungkinan seorang mahasiswa mengulang sebuah mata kuliah dipertimbangan pula dalam studi yang dilakukan.

Sebagian data yang diperoleh berisi informasi tersembunyi mengenai kinerja mahasiswa dan belum banyak dimanfaatkan untuk memperbaiki kualitas kinerja mahasiswa. Data ini digunakan untuk mempelajari data yang tersedia di bidang pendidikan dan membawa keluar pengetahuan tersembunyi yang ada pada data tersebut. Penelitian yang dilakukan untuk mendukung penyusunan skripsi. Berkaitan

e-ISSN: 2964-2531
DOI : 10.36499/psnst.v12i1.7012
p-ISSN: 2964-5131

dengan masalah yang muncul pada kemampuan akademik mahasiswa yang perlu di deteksi dimana bersifat menonjol yang cenderung positif atau cenderung negatif berpotensi anomali data akademik dapat menggambarkan kemampuan diatas atau dibawah rata – rata secara akademik. Dilihat pada sebagian mahasiswa mendapatkan nilai-nilai yang berbeda-beda dari setiap matakuliah yang diambil. Dengan perbedaan pada nilai tersebut perlu adanya identifikasi dalam permasalahan nilai untuk menangani prestasi nilai akademik nilai mahasiswa untuk meningkatkan proses pembelajaran di perlukan evaluasi yang mendalam terkait dengan kemampuan akademik dari setiap mahasiswa. Dengan demikian pengelolaan pengajaran lebih dapat di arahkan pada aspek – aspek penting yang berhubungan dengan pencapaian prestasi belajar. Untuk mengetahui kondisi akademik (Nilai) yang bersifat ekstrim baik positif atau negatif memerlukan kajian yang mendalam dari data akademik pada kurun waktu tertentu hal itu dengan data besar akan menimbulkan banyak kesulitan, karena semakin banyak jumlah data akademik akan menyebabkan kompleksitas yang bertambah. Perlu dilakukan suatu mekanisme yang dapat meng otomasi proses identifikasi data yang memiliki kecenderungan berbeda ekstrim di banding data lainnya. Diperlukan suatu sistem yang mampu mengidentifikasi melalui proses deteksi anomali.

## 2. METODELOGI

Pada bagian ini akan dijelaskan topik - topik tertentu berdasarkan review yang telah dilakukan terhadap beberapa literatur. Materi yang ditinjau dari literatur meliputi informasi tentang Algoritma DBSCAN, perhitungan mencari K-ideal, dan perbedaan *outlier* dengan *noise*.

### 2.1. Algoritma *DBSCAN*

Algoritma DBSCAN adalah salah satu algoritma *clustering density-based*. Sebagai metode deteksi berbasis klaster yang terkenal, dapat men-DBSCAN (Pengelompokan Aplikasi Spasial Berbasis Kepadatan dengan *Noise*). Algoritma *clustering* adalah salah satu metode yang paling utama untuk pengelompokan dalam penambangan data. DBSCAN memiliki kemampuan untuk menemukan kelompok ukuran dan bentuk variabel dan juga akan mendeteksi *noise*, dari sejumlah data besar yang mengandung *noise* dan *outlier*. DBSCAN akan mendeteksi klaster serta menentukan parameter epsilon secara otomatis dengan cara yang akurat untuk menemukan parameter input dan menemukan klaster dengan *outlier* yang berbeda-beda. Dengan demikian Eps, radius maksimum lingkungan, dan MinPts, jumlah minimum poin milik lingkungan Eps, adalah dua parameter *input* yang diperlukan untuk DBSCAN Algoritma memperluas wilayah dengan *outlier* yang tinggi ke dalam klaster dan menempatkan klaster irregular pada database spasial dengan *noise* (merepresentasikan *noise)* (Izhari, 2020).

Masalah mendeteksi klaster poin dalam data memang suatu hal yang menantang ketika klaster dari ukuran, kepadatan dan bentuk yang berbeda. Banyak masalah ini menjadi lebih signifikan ketika data dari dimensi yang sangat tinggi dan ketika itu termasuk ada *noise* dalam data dan *outlier*. Dalam hal ini algoritma DBSCAN clustering membutuhkan hanya satu parameter *input* dan mendukung pengguna dalam menentukan nilai yang sesuai untuk itu dan menemukan klaster sembarang bentuk. DBSCAN efisien bahkan untuk ruang yang besar database(goleman, daniel; boyatzis, Richard; Mckee, 2019). DBSCAN dapat mendeteksi kelompok yang berbagai bentuk dan ukuran dari sejumlah data besar yang mengandung *noise* dan *outlier*.

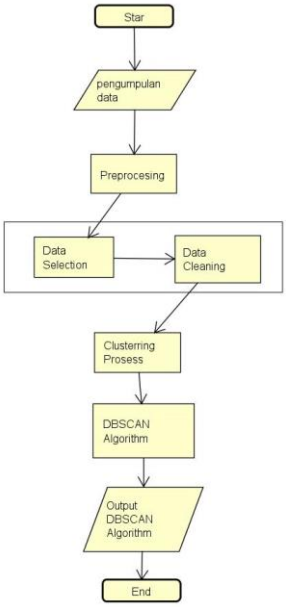
Metode ini mendefiniskan klaster sebagai maksimal set dari titik-titik yang *density-connected*. DBSCAN memiliki 2 parameter yaitu Eps (radius maksimum dari neighborhood) dan MinPts (jumlah minimum titik dalam Eps-neighborhood dari suatu titik). Untuk algoritma DBSCAN, dua parameter – dan MinPts – diperlukan (Xięski and Nowak-Brzezińska, 2017). Sebuah titik item dikatakan *density-reachable* dari titik item yang lain jika ada suatu rantai yang menghubungkan keduanya yang berisi hanya titik-titik yang *directly density-reachable* dari titik-titik sebelumnya. *Density-Based Methods* Merupakan metode yang dikembangkan berdasarkan *density* (kepadatan) tertentu. Metode ini menganggap klaster sebagai suatu area yang berisi objek-objek yang padat atau sesak, yang dipisahkan oleh area yang memiliki kepadatan rendah (merepresentasikan *noise*)(goleman, daniel; boyatzis, dkk 2019).

Pada Gambar 1, persiapan tersebut antara lain:

**1.** Data Selection
Data selection adalah Proses untuk memilih data dari database yang sesuai dengan tujuan analisis, proses meminimalkan jumlah data yang digunakan untuk proses mining dengan tetap mempresentasikan data aslinya.

**2.** Data Cleaning
Data Cleaning merupakan proses membersihkan data dari data tidak konsisten.Secara sederhana cara kerja DBSCAN adalah sebagai berikut :

1. Tentukan nilai minPts dan epsilon (eps) yang akan digunakan.
2. Pilih data awal "p" secara acak.
3. Hitung jarak antara data "p" terhadap semua data menggunakan Euclidian distance.
4. Ambil semua amatan yang density-reachable dengan amatan "p".
5. Jika amatan yang memenuhi nilai epsilon lebih dari jumlah minimal amatan dalam satu gerombol maka amatan "p" dikategorikan sebagai core points dan gerombol terbentuk.
6. Jika amatan "p" adalah border points dan tidak ada amatan yang density-reachable dengan amatan "p", maka lanjutkan pada amatan lainnya.
7. Ulangi langkah 3 sampai 6 hingga semua amatan diproses.




**Gambar 1. Flowchart DBSCAN Algorithm**

## 2.2. Euclidean Distance

*Euclidean Distance* atau jarak *Euclidean* adalah perhitungan jarak dari dua buah titik dalam *Euclidean space*. *Euclidean space* diperkenalkan oleh Euclid, Mari kita lihat contoh yang menunjukkan sifat non-deterministik dari algoritma DBSCAN. Katakanlah kita memiliki data dan parameter tertentu. Dengan rumus Euclidean Distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(Xi - Yi)}^{2} \quad (1)$$

Keterangan:
Xi : data uji
Yi : data nilai

e-ISSN: 2964-2531
DOI : 10.36499/psnst.v12i1.7012
p-ISSN: 2964-5131

### 2.3. Metode Elbow

Metode Elbow ini digunakan untuk menghasilkan informasi dalam menentukan jumlah klaster optimal dengan cara melihat persentase hasil perbandingan antara jumlah klaster yang akan membentuk siku pada suatu titik. dengan nilai k yang telah ditentukan dengan menggunakan metode elbow yang menunjukkan jumlah klaster terbaik (Aditya, Sari and Padilah, 2021). Metode Elbow digunakan untuk memilih jumlah klaster yang terbentuk pada suatu titik di grafik *Sum of Square Error* (SSE) dan didasarkan pada penurunan SSE yang besar. Jika nilai klaster sebelumnya (k-1) dengan nilai klaster selanjutnya (k) mengalami penurunan terbesar maka jumlah klaster tersebut yang tepat (k). Rumus SSE dapat dituliskan sebagai :

$$SSE=\sum\sum|xi-ck|^2XiKK=1 \quad (2)$$

Keterangan:
K: klaster. $xi$: data ke-i. $ck$: pusat klaster ke-k

### 2.4. Perbedaan Noise dan Outlier

DBSCAN cenderung memisahkan data yang mengandung *noise* agar tidak bercampur dengan klaster apapun yang ada dalam data, serta DBSCAN juga bekerja dengan memisahkan data berdasarkan kepadatan, dimana data padat terkonsentasi (jarak antar data dekat) versus data yang jarang atau berjarak jauh yang dikenali sebagai *noise* atau *outlier* (Ihsan Jambak and Efendi, 2021). Sekilas data *noise* dan *outlier* mungkin tampak sama. Namun, mereka benar - benar sangat berbeda. Oleh karena itu, Penting juga untuk mendefinisikan semua dasar yang berhubungan dengan *outlier*.

#### 2.4.1. Noise

DBSCAN mendefinisikan klaster sebagai himpunan maksimum dari titik-titik kepadatan yang terkoneksi (*density-connected*). Semua objek yang tidak masuk ke dalam klaster manapun dianggap sebagai *noise* (goleman, daniel; boyatzis,dkk 2019).

*Noise* yaitu data yang hanya membawa informasi yang tidak berarti. Berbagai masalah muncul dikehadiran data yang *noise* karena mesin tidak dapat memahami dengan benar dalam menafsirkannya, dengan penyebabnya(Smiti, 2020) :

1. Tipe data salah (jenis string dimasukkan untuk numerik atribut).
2. Nilai data yang salah (999 bukannya 99 untuk usia).
3. Nilai yang hilang (data tidak tercatat untuk suatu atribut).
4. Jenis data salah: (string dalam atribut numerik).
5. Data sangat berbeda dengan semua entri lainnya (10 dalam satu atribut jika tidak 0.1).

#### 2.4.2 Outlier

*Outlier* adalah titik data yang disebut pencilan. *Oulier* adalah juga disebut sebagai kelainan, discordants, menyimpang dan anomali. *Outlier* adalah konsep yang lebih luas yang mencakup tidak hanya kesalahan tetapi juga data sumbang yang mungkin timbul dari variasi alami dalam populasi atau proses. Dengan demikian, sering kali berisi informasi yang menarik dan berguna tentang sistem yang mendasarinya(Salgado *et al.*, 2016). Identifikasi outlier jika ada sedikit perbedaan pada hasil yang diperoleh maka *outlier* memiliki pengaruh yang minimal maupun maksimal, tetapi pengecualian mereka memang memiliki efek, itu mungkin lebih baik untuk menemukan alternatif. Di sinilah berbasis pengetahuan analisis *outlier* masuk (L.Sunitha, M.Bal Raju, 2013).

Kehadiran *outlier* dalam data dapat diartikan dengan :

1. Kesalahan pengukuran atau pencatatan ,
2. Nilai - nilai yang luar biasa tapi benar
3. Salah pelaporan

## 3. HASIL DAN PEMBAHASAN

### 3.1. Data Yang Dipakai

Proses pengambilan data nilai akademik diambil Data yang digunakan dalam penelitian ini adalah data akademik pada data mahasiswa informatika tahun 2014 - 2017 yang diperoleh pada Universitas

Jenderal Achmad Yani. Data akademik yang digunakan sebanyak 200 data. Data tersebut terdiri dari semeseter 1 sampai semester 6 dari hasil nilai ujian akhir semester (UAS) dan nilai akhir yang berupa sebuah dokumen excel. Dengan Id. Yang mana mata kuliah inti informatika yang diambil dengan nilai Ulangan Akhir Semester (UAS) dan nilai akhir (NH). Yaitu Algoritma & Pemprograman, Praktikum Algoritma & Pemprograman, Pemprograman Objek 1, Praktikum Pemprograman Objek 1, Basis Data, Praktikum Basis Data, Data Warehouse Dan Data Mining, Implementasi Perangkat Lunak, Praktikum Implementasi Perangkat Lunak, Pemprograman Web, Praktikum Pemprograman Web, Pemprograman Objek 2, Praktikum Pemprograman Objek 2, Rekayasa Perangkat Lunak, Kecerdasan Buatan, Prak Kecerdasan Buatan, Teknologi Web, Praktikum Teknologi Web, Mobile Programming, Praktikum Mobile Programming.

### 3.2. Perhitungan Data Yang Dipakai

**Tabel 1 Dataset pemisalan**

| Id | IF1141 | | IF1142 | | IF2125 | | IF2126 | |
|---|---|---|---|---|---|---|---|---|
| | UAS | NH | UAS | NH | UAS | NH | UAS | NH |
| 1 | 45 | 62 | 54 | 58 | 56 | 57 | 56 | 56 |
| 2 | 77 | 89 | 76 | 80 | 65 | 76 | 56 | 62 |
| 3 | 67 | 86 | 77 | 81 | 79 | 80 | 79 | 80 |
| 4 | 70 | 67 | 69 | 68 | 68 | 68 | 68 | 68 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 200 | 55 | 57 | 76 | 78 | 78 | 80 | 78 | 80 |

#### 3.2.1. Perhitungan Euclidean Distance

Hasil pembobotan data yang sudah dilakukan kemudian digunakan dalam proses density-based clustering. Algoritma DBSCAN yang akan diimplementasikan akan membuat klaster sesuai dengan parameter masukan, yaitu $\epsilon$ dan MinPts. Parameter $\epsilon$ dan MinPts akan mempengaruhi jumlah klaster yang terbentuk. DBSCAN akan membuat suatu daerah yang berpusat di $\theta$ dengan radius sebesar $\epsilon$, sehingga anggota klaster adalah objek-objek dalam radius $\epsilon$ dari objek pusat $\theta$. Perhitungan jarak objek p ke objek pusat $\theta$ dapat menggunakan pengukuran numerik yaitu menggunakan Euclidean Distance. Berikut rumus Euclidean Distance pada rumus persamaan --- (1)

**Tabel 2 Pemisalan Euclidean DIstance**

| UAS | NH | UAS | NH | UAS | NH | |
|---|---|---|---|---|---|---|
| IF1141 | | IF1142 | | IF2125 | | |
| 50 | 50 | 50 | 50 | 50 | 50 | IDuji |
| 45 | 62 | 54 | 58 | 56 | 57 | ID1 |
| 77 | 89 | 76 | 80 | 65 | 76 | ID2 |
| 67 | 86 | 77 | 81 | 79 | 80 | ID3 |

Diketahui Xi, X1= 50, X2 = 50, X3 = 50, X4 = 50, X5 = 50, X6 = 50 maka jika dihitung menggunakan rumus Eclidean distance, maka didapatkan hasil perhitungan jarak Euclidean Distance sebagai tabel berikut. Hasil perhitungan ini diurutkan dari jarak yang terbesar hingga terkecil pada tabel 3.

e-ISSN: 2964-2531
p-ISSN: 2964-5131

**Tabel 3 Hasil Euclidean Distance**

| Nama | *Euclidean Distance* |
|---|---|
| D2 | 70,82 |
| D3 | 68,75 |
| D1 | 18,27 |

Dilihat bahwa matrik jarak ini merupakan matrik simetri dengan nilai dengan dan diagonalnya bernilai 0 (nol). Dikarenakan matrik jarak ini merupakan tolak ukur kemiripan yang digunakan untuk menggelompokkan data maka semakin kecil jarak *euclidean* akan semakin mirip pula kedua objek tersebut dan nilai 0 pada diagonal berarti bahwa ada jarak antara objek atau dikatakan sangat mirip.

### 3.2.2 Perhitungan Metode Elbow

Perhitungan Elbow dalam menentukan klaster ideal, pada rumus persamaan (2) dengan hasil sebagai berikut :

Data pengamatan = 3
Xi = 18,27 ; 68,75 ; 70,82
ck = 52,6
(Xi – ck ) = -34,33 ; 16,15 ; 18,220
$(Xi – ck)^2$ = 1.178,5 ; 260,8 ; 331,9
$\Sigma(Xi – ck)^2 = 1.771,2 \approx 2$

### 3.3. Hasil Pengujian dan Analisa Hasil

Yang didapat dari eksperimen diatas menunjukkan bahwa dataset nilai terdapat klaster ideal berjumlah 2 yang diberikan, dengan hasil DBSCAN yang diusulkan, dan telah tercatat bahwa performa terdeteksinya *outlier* sebesar 5% saat ini untuk dataset yang ada dan bagi mahasiswa yang terdeteksi adanya anomali dapat segera dilakukannya evaluasi terutama pada minimum *outlier*.

## 4. KESIMPULAN

Dari penelitian sebelumnya dengan menggunakan metode yang berbeda, pada penelitian di atas diperoleh algoritma dengan tingkat akurat yang lebih baik dan dapat disimpulkan bahwa pada hasil penelitian ini terdapat 2 klaster pada dataset sebanyak 200 data nilai, dimana hasil metode DBSCAN mendeteksi *outlier* sebesar 5%, dimana 5% itu merupakan titik yang terjauh dari jarak titik centroid dan tidak termasuk ke dalam *cluster* manapun. Serta DBSCAN mampu mengidentifikasi anomali secara baik. Baik itu nilai anomali yang tinggi ataupun anomali yang rendah. Hal ini menggambarkan sedikitnya anomali data, dengan demikian upaya – upaya penanganan terhadap kemampuan akademik mahasiswa sebesar 5% di rekomendasikan untuk ditangani secara khusus. Dan pada dataset ini di simpulkan bahwa DBSCAN juga bisa tidak mempresentasikan data spasial.

## DAFTAR PUSTAKA

Aditya, A., Sari, B. N. and Padilah, T. N. (2021) ‘Comparison analysis of Euclidean and Gower distance measures on k-medoids cluster’, *Jurnal Teknologi dan Sistem Komputer*, 9(1), pp. 1–7. doi: 10.14710/jtsiskom.2020.13747.

goleman, daniel; boyatzis, Richard; Mckee, A. (2019) ‘Algoritma DBSCAN dan Contoh Perhitungannya’, *Journal of Chemical Information and Modeling*, 53(9), pp. 1689–1699.

Ihsan Jambak, M. and Efendi, R. (2021) ‘Pengaruh Reduksi Dimensi Terhadap Metode Pengklasteran Berbasis Centroid dan Metode Pengklasteran Berbasis Density dalam Pengklasteran Dokumen Teks’, *Indonesian Journal of Business Intelligence*, 4(2), pp. 53–62. Available at: http://dx.doi.org/10.21927/ijubi.v4i2.1918.

Izhari, F. (2020) ‘Analisis Algoritma Dbscan Dalam Menentukan Parameter Epsilon Pada Clustering

Data Numerik', *Seminar Nasional Teknologi Komputer & Sains*, pp. 156–158.
L.Sunitha, M.Bal Raju, B. S. S. (2013) 'A Comparative Study between Noisy Data and Outlier Data in Data Mining', *International Journal of Current Engineering and Technology*, 3(2), pp. 575–577. Available at: http://inpressco.com/wp-content/uploads/2013/06/Paper64575-577.pdf.
Mayadewi, P. and Rosely, E. (2015) 'Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining', *Seminar Nasional Sistem Informasi Indonesia*, (November), pp. 329–334.
Novalia, V. and Goejantoro, R. (2020) 'Perbandingan Metode Klasifikasi Naive Bayes dan K-Nearest Neighbor The Comparison Method Of Classification Naive Bayes and K-Nearest Neighbor ( Case Study : Employment Status Of Citizen In Kutai Kartanegara Regency 2018 )', 11, pp. 159–166.
Salgado, C. M. *et al.* (2016) 'Secondary Analysis of Electronic Health Records', *Secondary Analysis of Electronic Health Records*, pp. 1–427. doi: 10.1007/978-3-319-43742-2.
Smiti, A. (2020) 'A critical overview of outlier detection methods', *Computer Science Review*. Elsevier Inc., 38, p. 100306. doi: 10.1016/j.cosrev.2020.100306.
Xięski, T. and Nowak-Brzezińska, A. (2017) 'Outlier mining using the DBSCAN algorithm', *Journal of Applied Computer Science*, 25(January 2017), pp. 53–68.