Outlier Detection On Graduation Data Of Darussalam Gontor University Using One-Class Support Vector Machine

Deteksi Outlier Pada Data Kelulusan Mahasiswa Universitas Darussalam Gontor Menggunakan One-Class Support Vector Machine

Oddy Virgantara Putra, Triana Harmini, Ahmad Saroji
{oddy@unida.gontor.ac.id[1] , triana@unida.gontor.ac.id[2], ahmadsaroji@mhs.unida.gontor.ac.id[3] }

Universitas Darussalam Gontor

Abstract*. Outlier detection is an important field of study because it is able to detect abnormal data distribution from a set of data. In the student graduation data, there are students with high semester GPA but do not graduate on time but students with low semester GPA can graduate on time. This study aims to detect outlier values in student graduation data for the 2020-2021 class. Factors (attributes) used in this study are Student Academic Support Credit Scores (AKPAM) and Social Studies from semester one to semester six. The dataset used is 1204 graduates. The outlier detection method used is One-Class Support Vector Machine (SVM). One-class SVM is a derivative of SVM method that detects outliers based on data outside the specified class. The results of outlier detection using One-Class SVM method with three types of kernels produce the following reference values: kernel 'rbf' n by 91.4%, kernel 'linear' by 90% and kernel 'poly' by 90%. After normalization using MinMaxScaler the reference value increased by 2% in each kernel. The results of testing the One-Class SVM method get an average 90.3%, thus it can be concluded that the One-Class SVM method is feasible to be used as an outlier detection method..*

Keywords *- Student graduation; outlier detection; SVM; One-Class SVM; MinMaxScaler.*

Abstrak*. Deteksi Outlier merupakan salah satu bidang kajian yang penting karena mampu mendeteksi sebaran data tidak normal dari sekumpulan data. Pada data kelulusan mahasiswa terdapat mahasiswa dengan IP semester tinggi namun tidak lulus tepat waktu tetapi mahasiswa dengan IP semester rendah dapat lulus tepat waktu. Penelitian ini bertujuan untuk mendeteksi nilai outlier pada data kelulusan mahasiswa angkatan 2020-2021. Faktor(atribut) yang digunakan dalam penelitian ini adalah Angka Kredit Penunjang Akademik Mahasiswa (AKPAM) dan IPS dari semester satu sampai semester enam. Dataset yang digunakan sebanyak 1204 lulusan. Metode deteksi outlier yang digunakan adalah One-Class Support Vector Machine (SVM). One-class SVM merupakan turunan dari metode SVM yang mendeteksi outlier berdasarkan data diluar kelas yang telah ditentukan. Hasil dari outlier detection menggunakan metode One-Class SVM dengan tiga jenis kernel menghasilkan nilai acuan sebagai berikut: kernel 'rbf' n sebesar 91,4%, kernel 'linear' sebesar 90% dan kernel 'poly' sebesar 90%. Setelah dilakukan normalisasi mengggunakan MinMaxScaler nilai acuan meningkat sebesar 2% di setiap kernel. Hasil dari pengujian metode One-Class SVM mendapatkan nilai rata-rata sebesar 90,3%, dengan demikian dapat disimpulkan bahwa metode One-Class SVM layak untuk digunakan sebagai metode deteksi outlier.*

Kata Kunci *- Kelulusan mahasiswa; deteksi outlier; SVM; One-Class SVM; MinMaxScaler*

I. PENDAHULUAN

Universitas Darussalam (UNIDA) Gontor merupakan perguruan tinggi pesantren yang terletak di Ponorogo Jawa Timur. UNIDA Gontor memiliki jumlah mahasiswa dan mahasiswi sebanyak 5237dengan rincian sebagai berikut; sebanyak 49% mahasiswa dan 51% mahasiswi, sebanyak 40% mahasiswa/i guru dan mahasiswa/i murni 60%. UNIDA Gontor memiliki program pendidikan di 7 fakultas dan 17 program studi sarjana (S1), 3 program studi pascasarjana (S2), dan 1 program doctoral (S3). Data ini diambil pada hari Kamis, 23 September 2021. [1]

Berdasarkan data kelulusan mahasiswa pada tahun 2019 dan 2020, sebanyak 19% dan 15% tidak lulus tepat waktu. Kelulusan mahasiswa dipengaruhi oleh dua faktor yaitu nilai Angka Kredit Penunjang Akademik Mahasiswa (AKPAM) dan Nilai IPS. Pada data kelulusan mahasiswa terdapat mahasiswa yang mendapat nilai IPS dan nilai AKPAM rendah akan tetapi mahasiswa tersebut bisa menyelesaikan tepat waktu, sedangkan terdapat mahasiswa yang mendapat nilai IPS dan AKPAM tinggi akan tetapi mahasiswa tersebut tidak bisa menyelesaikan studinya tepat waktu. Pada data kelulusan mahasiswa terdapat mahasiswa lulus tepat waktu dan lulus tidak tepat waktu. Data mahasiswa tidak tepat waktu tersebut merupakan data *anomaly*, dikarenakan data tersebut data yang berbeda dengan data normal yaitu data mahasiswa lulus tepat waktu.

Dalam memprediksi kelulusan mahasiswa dengan data *imbalance* dengan perbandingan 89% data mahasiswa lulus tepat waktu dan 11% data mahasiswa tidak lulus tepat waktu, dari keseluruhan jumlah dataset. Maka data tersebut

tidak bisa digunakan untuk data klasifikasi. Maka dari itu, diperlukanlah metode untuk mendeteksi *outlier* terhadap data *anomaly.*

Pada penelitian pertama yang dilakukan pada tahun 2018 oleh Tung Kieu, Bin Yang, Christian S. Jensen. Metode yang mereka gunakan adalah *One-Class SVM* dan *Local Outlier Factor* dengan mendapatkan nilai acuan atau *F1-Score* dengan metode *One-Class SVM* sebesar 67,3% dan *Local Outlier Factor* 97,4%[2]. Kemudian penelitian kedua yang dilakukan pada tahun 2020 oleh Canghong Shi, Xiaojie Li, Jiancheng Lv, Jing Yin, dan Imran Mumtaz dengan menggunakan tiga metode yaitu *One-Class SVM, Isolation Forest, dan Local Outlier Factor.* Ketiga metode tersebut menghasilkan dengan masing-masing nilai akurasi *One-Class SVM 82%, Islation Forest 98%, dan Local Outlier Factor 98%.*[3]

Ada beberapa metode yang dapat digunakan untuk deteksi *outlier*, salah satunya adalah metode *One-Class Support Vector Machine (OCSVM).* Metode *OCSVM* adalah sebuah metode turunan dari metode *Support Vector Machine* yang berfungsi untuk mendeteksi data yang berada di luar kelas yang telah ditentukan yaitu *outlier.* Tujuan dari penelitian ini adalah mengetahui berapa jumlah data yang terdeteksi benar menggunakan metode *OCSVM.*

II. METODE

Pada penelitian ini terdapat langkah-langkah yang dilakukan dalam penelitian. Langkah-langkah yang dilakukan adalah pengambilan dataset, dimana dalam pengambilan dataset ada dua tahapan yaitu: pengumpulan data dan pelabelan data. Langkah selanjutnya adalah *pre-processing*, didalam tahap *pre-processing* ada dua tahap yaitu: *missing data* dan *minmaxscaller*. Setelah melalui tahap *pre-processing* maka dataset yang akan digunakan sudah siap untuk diolah, yaitu dengan menggunakan metode *OCSVM*. Setelah data diolah menggunakan metode *OCSVM*, langkah selanjutnya adalah evaluasi. Tahap evaluasi bertujuan untuk mengetahui kinerja suatu metode terhadap data yang diolah, dengan demikian dapat diketahui apakah metode ini layak atau tidak untuk digunakan.

A. Pengambilan dataset

Pada tahap pengumpulan dataset terdapat 2 langkah yang dilakukan, yaitu pengumpulan data dan pelabelan data.

Pengumpulan data

Dalam penelitian ini ada dua faktor(atribut) yang digunakan untuk mendeteksi outlier pada data kelulusan mahasiswa yaitu nilai Angka Kredit Penunjang Akademik Mahasiswa dan nilai IPs dari semester satu sampai enam. Pengumpulan data diambil dari staff Biro Akreditasi Akademik Kemahasiswaan (BAAK) untuk mengambil data kelulusan mahsiswa pada angkatan 35,36 dan 37. Kemudian mengambil data IPs dari semester satu sampai semester enam. Kemudian pengambilan data AKPAM kepada Direktorat Kepesantrenan (DKP). Rentang nilai IPs adalah dari nol sampai empat, sedangkan rentang nilai AKPAM dari 200 sampai 400. Setelah data telah didapatkan, langkah selanjutnya adalah mengisi dan mencocokan data kelulusan mahasiswa dengan nilai AKPAM dan IPs disetiap semesternya hingga semester enam.

Pelabelan data

Pelabelan data dilakukan secara manual yaitu dengan melihat tahun masuk mahasiswa dan tahun kelulusan mahasiswa. Jika mahasiswa tersebut bisa menyelesaikan studinya selama empat tahun dari jarak tahun masuk mahasiswa tersebut maka mahasiswa tersebut tepat waktu. Sedangkan mahasiswa yang menyelesaikan studinya lebih dari empat tahun dari jarak tahun masuk, maka mahasiswa tersebut tidak tepat waktu. Mahasiswa lulus tepat waktu diberikan label (1) dan mahasiswa lulus tidak tepat waktu diberikan label (0).

B. Preprocessing

Pre-processing data merupakan tahap awal untuk menyiapkan data yang telah diambil agar siap untuk di proses. Proses ini juga bisa disebut sebagai proses untuk mengubah data menjadi lebih bersih dan menggabungkan data tersebut untuk proses selanjutnya. Pada penelitian ini terdapat 2 tahap dalam *pre-processing* data.

Missing data

Missing data merupakan hilangnya informasi atau tidak tersedianya suatu data dalam sebuah obyek. Missing data merupakan masalah yang sering dijumpai dalam sebuah penelitian *data mining.* Keberadaan *data missing* dapat menggangu dalam sebuah penelitian. Pada penelitian ini terdapat 248 kolom *data missing* dari total keseluruhan kolom sebanyak 14448, maka dengan demikian jumlah presentase *data missing* sebanyak 2% dari jumlah data. Jika jumlah *data missing* berjumlah 1- 5%, maka data tersebut masih bisa diolah[4]. Pengisian *data missing* menggunakan *function library* pandas yaitu mengisi data dengan cara menduplicate data sebelumnya.

Minmaxscaller

Pada tahap *pre-processing* terdapat tahap normalisasi data. Normalisasi adalah proses penskalaan nilai atribut dari data sehingga data tersebut dapat terletak pada rentang tertentu. Dalam normalisasi data terdapat beberapa teknik yang digunakan untuk normalisasi data, salah satunya adalah *MinMaxScaller*. *MinMaxScaller* merupakan metode

normalisasi dengan melakukan transformasi linier pada data asli. Sehingga dapat menghasilkan keseimbangan nilai perbandingan antar data[5]. Perhitungan *MinMaxScaller* dapat menggunakan rumus sebagai berikut:

$$Xsc = \frac{X - Xmin}{Xmax - Xmin}$$

Xsc = Nilai normalisasi.
X = Nilai *value* dalam dataset.
Xmin = Nilai minimal dalam suatu kolom.
Xmax = Nilai maximal dalam suatu kolom.

.

C. Data Processing

Pada tahap ini mulai diterapkanya algoritma *one-class svm* yang merupakan salah satu metode yag digunakan untuk mendeteksi data pencilan atau *outlier.* Metode *one-class svm* mempunyai bebrapa kernel yang bisa digunakan dalam *outlier.*Pada penelitian ini menggunakan tiga kernel yaitu RBF (*Radial Basis Function*),*Linear* dan *Polynomial*.[6]

D. Evaluasi

Dalam uji coba diperlukannya sebuah evaluasi untuk mengetahui suatu metode yang digunakan. Kinerja suatu metode dapat dievaluasi dengan menghitung nilai *True Positive, False Negative, False Positive, dan True Negative. True Positive* adalah sampel positif yang dideteksi sebagai positif oleh *classifier.True Negative* adalah sampel negative yang benar dan terdeteksi negatif . *False Positive* adalah sampel negative yang salah dan terdeteksi sebagai positif. *False Negative* adalah sample positif yang salah dan terdeteksi negative.

F1-Score merupakan perbandingan nilai rata-rata presisi dan recall yang dibobotkan, dengan rumus sebagai berikut:

$$F1\text{-}Score = 2 \text{ x } \frac{Recall\ x\ Presisi}{Recall + presisi}$$

F1-Score: perbandingan rata-rata presisi dan recall yang dibobotkan.
Recall: rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.
Presisi: rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positf.

Pada penelitin ini sebagai evaluasi untuk mengetahui kinerja dari suatu metode, maka digunakanlah nilai *F1-Score* yang dijadikan sebagai nilai acuan peforma metode tersebut. *F1-Score* adalah nilai acuan yang diginakan ketika dataset yang digunakan dalam perhitungan *imbalance.* Karena data *imbalance* akan menyebabkan jumlah data *False Negative* dan *FalsePositive* tidak *Symmetric.* Nilai akurasi hanya bisa digunakan sebagai nilai acuan jika data tersebut *balance*[7].

III. HASIL DAN PEMBAHASAN

Dalam penelitian ini data yang digunakan adalah data kelulusan mahasiswa pada tahun 2020-202. Jumlah daya kelulusan mahasiswa sebanyak 1204 yang terdiri dari 3 angkatan yaitu, 35,36 dan 37. Data dibagi menjadi 2 untuk dijadikan sebagai data *train* dan data *test.* Data di bagi dengan skala **80 x 20** secara random , sehingga diperoleh data *train* **963 data** dan data *test* berjumlah **241 data.**Dalam penelitian ini menggunakan laptop Acer Nitro AN515-43 dengan AMD Ryzen 5 3550H *with* Radeon Vega Mobile Gfx (8CPUs),~2.1GHz, *Random Acces Memory* (RAM) 8 GB DDR 4, *64-bit operating system windows 10 pro.*

A. Hasil model one class svm dengan kernel rbf(radial basis function)

Hasil pengujian dengan menggunakan metode *one-class svm* dengan kernel RBF mendapatkan nilai F1-Score sebesar 91,4%. Nilai tersebut disebut sebagai nilai acuan, yaitu nilai yang diambil dari rata-rata nilai presisi dan nilai recall. Dari nilai F1-Score dapat diketahui bahwa metode tersebut layak atau tidak.

B. Hasil model one class svm dengan kernel linear

Hasil pengujian dengan menggunakan metode *one-class svm* dengan kernel linear mendapatkan nilai F1-Score sebesar 90,0%. Nilai tersebut disebut sebagai nilai acuan, yaitu nilai yang diambil dari rata-rata nilai presisi dan nilai recall. Dari nilai F1-Score tersebut dapat diketahui bahwa metode tersebut layak atau tidak.

C. Hasil Model One Class SVM Dengan Kernel Polynomial

Hasil pengujian dengan menggunakan metode *one-class svm* dengan kernel polynomial mendapatkan nilai F1-Score sebesar 90,0%. Nilai tersebut disebut sebagai nilai acuan, yaitu nilai yang diambil dari rata-rata nilai presisi dan nilai recall. Dari nilai F1-Score tersebut dapat diketahui bahwa metode tersebut layak atau tidak.

Setelah selesai melalui tahap pengujian mengggunakan tiga kernel dari metode *one-class svm* langkah selanjutnya dilakukan tahap *minmaxscaller.* Fungsi dari *minmaxscaller* adalah untuk menormalisasi data untuk mengetahui jarak antar data[8]. Setelah dilakukan normalisasi pada data nilai *F1-Score* atau nilai acuan mengalami peningkatan sebesar 2% pada setiap kernel. Peningkatan tersebut dapat dipresentasikan pada **Table 1.**

Table 1. Hasil *F1-Score* setelah normalisasi dengan *MinMaxScaller*

One-Class SVM	Tanpa MinMaxScaller	MinMaxScaller
Kernel RBF	91.4%	93.4%
Kernel Linear	90.0%	92.0%
Kernel Polynomial	90.0%	92.0%

VII. KESIMPULAN

Berdasarkan dari hasil penelitian yang telah dilakukan pada penelitian ini, dapat disimpulkan bahwa deteksi *outlier* dengan menggunakan metode *One-Class SVM* dengan menggunakan tiga kernel menghasilkan nilai *F1-Score* sebagai berikut; kernel RBF sebesar 91.4%, kernel *linear* 90.0% dan kernel polynomial 90.0%. Nilai *F1-Score* tersebut merupakan nilai hail dari perhitungan data sebelum melalui normalisasi. Setelah dilakukan normalisasi pada data, nilai *F1-Score* mengalami peningkatan sebesar 2%. Hasilnya sebagai berikut; kernel RBF sebesar 93.4%, kernel *linear* dan polynomial sebesar 92.0%. Normalisasi data berpengaruh terhadap hasil kinerja pada suatu metode. Maka dengan demikian metode *OCSVM* layak untuk digunakan untuk mendeteksi *outlier* dalam suatu data.

UCAPAN TERIMA KASIH

Penelitian ini didanai oleh Program Studi Teknik Informatika Universitas Darussalam Gontor Ponorogo Indonesia. Ucapan terima kasih diberikan kepada para dosen Program Studi Teknik Informatika yang telah membimbing dengan ikhlas dan sepenuh hati.

REVERENSI

[1] I. Kartikasari, “laporan kamisan 23 September 2021.pdf,” BAAK Data, Ponorogo.
[2] T. Kieu, B. Yang, and C. S. Jensen, “Outlier detection for multidimensional time series using deep neural networks,” *Proc. - IEEE Int. Conf. Mob. Data Manag.*, vol. 2018-June, pp. 125–134, 2018, doi: 10.1109/MDM.2018.00029.
[3] P. Nair and I. Kashyap, “Hybrid Pre-processing Technique for Handling Imbalanced Data and Detecting Outliers for KNN Classifier,” *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 460–464, 2019, doi: 10.1109/COMITCon.2019.8862250.
[4] I. Atastina, “Analysis of missing value handling by collateral missing value estimation (cmve) method,” 2011.
[5] D. A. Nasution, H. H. Khotimah, and N. Chamidah, “Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN,” *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.
[6] Y. Wang, J. Wong, and A. Miner, “Anomaly intrusion detection using one class SVM,” *Proc. fron Fifth Annu. IEEE Syst. Man Cybern. Inf. Assur. Work. SMC*, pp. 358–364, 2004, doi: 10.1109/iaw.2004.1437839.
[7] R. Arthana, “Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning,” *Arthana, Resika*. https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8.
[8] D. S. Informasi, *MENGGUNAKAN METODE SUPPORT VECTOR MACHINE FORECASTING THE NUMBER OF TUBERCULOSIS DISEASE PATIENTS IN EAST JAVA REGION USING*. 2019.