Jurnal Smart Teknologi
Vol. 3, No. 3, Maret 2022, Halaman 261 – 266
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

Analisis Perbandingan Metode *K Nearest Neighbor* Dan *Gaussian Naive Bayes* Pada Klasifikasi Jurusan Siswa (Studi Kasus pada Siswa SMA Muhammadiyah 3 Jember)

A Comparative Analysis of K Nearest Neighbor and Gaussian Naive Bayes methods in students major classification (A Case Study Of SMA Muhammadiyah 3 Jember)

Herdian Cahyaningrum[1], Deni Arifianto[2*], Ginanjar Abdurrahman[3]
[1]Mahasiswa Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember
email: dyench.herdian@gmail.com
[2]Dosen Program Studi Teknik Informatika, Fakultas Teknik, *Koresponden Author
email: deniarifianto@unmuhjember.ac.id
[3]Dosen Program Studi Teknik Informatika, Fakultas Teknik,
email: abdurrahmanginanjar@unmuhjember.ac.id

ABSTRAK

Penelitian ini akan mengupas tentang klasifikasi penjurusan siswa menggunakan metode *Gaussian Naive Bayes* dan *K Nearest Neighbor*. Penelitian ini menggunakan data siswa SMA Muhammadiyah 3 dengan fitur yang digunakan dalam penjurusan siswa adalah rekapitulasi nilai ujian nasional bahasa Indonesia, nilai ujian nasional bahasa inggris, nilai ujian nasional IPA, nilai ujian nasional matematika, ulangan harian IPA, Matematika, IPS, Bahasa Indonesia, Bahasa inggris, nilai tes verbal linguistik, logis matematis, spasial, kinestik, musikal, interpersonal, intrapersonal dan natural. Total data yang dihitung berjumlah 320 data. *Preprocessing* data menggunakan metode *Median Subtitution* dan metode *Min-Max Normalization*. Untuk mengatasi ketidakseimbangan data pada penelitian ini menggunakan metode *SMOTE (Synthetic Minority Oversampling Technique)*. Dari data sintetis hasil *SMOTE* diperoleh data total 486 data. Skenario uji dalam penelitian ini menggunakan metode *K Fold Cross Validation* dengan nilai *k Fold* = 2, 4, 5, 8 dan 10. Dalam pengukuran jarak, vektor yang digunakan dalam implementasi *K Nearest Neighbor* menggunakan *Euclidean Distance*. Hasil akurasi tertinggi metode *Gaussian Naive Bayes* adalah 83,33% sedangkan akurasi tertinggi yang diperoleh metode *K Nearest Neighbor* adalah 83,61% dengan nilai *k neighbor* = 7.

Kata Kunci: *Gaussian Naive Bayes*; Klasifikasi; *K-NN*; *SMOTE*

ABSTRACT

This study will explore the classification of students' majors using the Gaussian Naive Bayes and K Nearest Neighbor methods. This study uses data from SMA Muhammadiyah 3 students with the features used in student majors are recapitulation of Indonesian national exam scores, English national exam scores, science national exam scores, math national exam scores, science daily tests, mathematics, social studies, Indonesian language, English, verbal linguistics, logical mathematical, spatial, kinesthetic, musical, interpersonal, intrapersonal and natural test scores. The total calculated data is 320 data. Preprocessing data using the Median Substitution method and the Min-Max Normalization method. To overcome the imbalance of data in this study using the SMOTE (Synthetic Minority Oversampling Technique) method. From the synthetic data, the results of SMOTE obtained a total of 486 data. The test scenario in this study uses the K Fold Cross Validation method with k Fold values = 2, 4, 5, 8 and 10. In measuring distance, the vector used in the implementation of K Nearest Neighbor uses Euclidean Distance. The highest accuracy result of the Gaussian Naive Bayes method is 83.33% while the highest accuracy obtained by the K Nearest Neighbor method is 83.61% with the value of k neighbor = 7.

Jurnal Smart Teknologi
Vol. 3, No. 3, Maret 2022, Halaman 261 – 266
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

Keyword : Gaussian Naive Bayes; Classification; K-NN; SMOTE.

1. PENDAHULUAN

SMA Muhammadiyah 3 Jember merupakan sekolah swasta yang memiliki tiga penjurusan yaitu "bahasa", "ipa" dan "ips". Dalam penentuan jurusan pada siswa, SMA Muhammadiyah menggunakan rekapitulasi nilai ujian nasional bahasa Indonesia, nilai ujian nasional bahasa inggris, nilai ujian nasional IPA, nilai ujian nasional matematika, ulangan harian IPA, Matematika, IPS, Bahasa Indonesia, Bahasa inggris, nilai tes verbal linguistik, logis matematis, spasial, kinestik, musikal, interpersonal, intrapersonal dan natural [1]. Berbagai permasalahan tentang data siswa dalam melakukan penjurusan dihadapi pihak sekolah. Seperti tidak mengikuti tes, sehingga banyak data siswa tanpa nilai pada fitur tertentu. Penelitian ini akan berfokus pada cara mengatasi *Missing value* pada data siswa serta mengatasi ketidakseimbangan data pada data siswa.

Penelitian ini memanfaatkan konsep *Data Mining* dalam mengatasi permasalahan penjurusan. *Data Mining* merupakan teknik untuk menggali informasi serta pengetahuan yang tersimpan dalam sebuah data menggunakan statistik, matematika serta *Machine Learning* [2]. Teknik *Data Mining* yang digunakan dalam penelitian ini adalah klasifikasi. Metode klasifikasi pada *Data Mining* bermacam-macam, diantaranya adalah *Gaussian Naive Bayes* dan *K Nearest Neighbor*.

Gaussian Naive Bayes merupakan salah satu varian dari Naive Bayes, di mana metode ini bekerja melalui probabilitas. Pada metode ini data yang digunakan pada fitur yang digunakan adalah tipe angka. Metode klasifikasi lainnya yaitu *K Nearest Neighbor*, merupakan metode klasifikasi yang dapat bekerja pada data angka. Metode ini bekerja dengan cara mengukur jarak terdekat dalam mengklasifikasi sebuah data [3].

Dikarenakan metode *Gaussian Naive Bayes* dan *K Nearest Neighbor* memiliki karakter yang sama, yaitu dapat bekerja pada data angka, maka dalam penelitian ini akan dilakukan perbandingan performa dari metode keduanya. Pengukuran performa dilakukan pada tingkat akurasi, presisi dan *recall* metode keduanya. Total data yang digunakan dalam penelitian ini sebanyak 320 data siswa. Dalam pengukuran jarak pada metode *K Nearest Neighbor*, menggunakan *Euclidean Distance* dengan nilai ketetanggaan k = 3, 5, 7 dan 9. Pemodelan pada penelitian ini menggunakan skenario uji *K Fold Cross Validation*. Pada *preprocessing* data akan digunakan metode *Median Subtitution* dan *Min-Max Normalization* serta penggunaan *SMOTE (Synthetic Minority Oversampling Technique)* dalam mengatasi ketidakseimbangan data.

2. KAJIAN PUSTAKA

Penelitian ini tidak terlepas dari literatur yang digunakan sebagai landasan, dasar serta pedoman dalam menyelesaikan penelitian ini.

A. *Missing value*

Missing value merupakan hilangnya suatu informasi atau tidak tersedianya informasi/objek tertentu pada suatu variabel dalam sebuah data yang diakibatkan dari faktor tertentu [4]. Pada penelitian ini digunakan *Median Subtitution* dalam mengatasi *Missing value*.

$$data\ ganjil, Me = x_{\frac{1}{2}(n+1)} \quad (1)$$

$$data\ genap, Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \quad (2)$$

B. Normalisasi

Normalisasi ini dapat berfungsi untuk mempermudah pembandingan nilai dalam sebuah data yang memiliki ukuran angka yang berbeda [5]. Berikut persamaan *Min-Max Normalization*.

$$z = \frac{X - \min(x)}{[\max(x) - \min(x)]} \quad (3)$$

dimana, x merupakan atribut, sedangkan X merupakan data ke i pada sebuah data.

C. *SMOTE (Synthetic Minority Oversampling Technique)*

Metode *SMOTE (Synthetic Minority Oversampling Technique)* bekerja dengan menambah jumlah data minor sehingga setara

Jurnal Smart Teknologi
Vol. 3, No. 3, Maret 2022, Halaman 261 – 266
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

dengan jumlah data kelas mayor. Cara yang dilakukan adalah dengan membangkitkan atau membuat data buatan (sintetis) [6].

$$Euclidean\ Distance = \sqrt{\sum_{k=1}^{n}(X_k - Y_k)^2} \quad (4)$$

Dimana, X merupakan data yang ingin dikloning, sedangkan Y merupakan data yang dibangkitkan untuk memperoleh data baru.

D. *K Fold Cross Validation*

K Fold Cross Validation adalah sebuah teknik untuk menemukan performa terbaik dalam sebuah metode. Teknik ini bekerja dengan cara membagi data menjadi beberapa lipatan k. Tiap lipatan k akan memiliki porsi sama besar [7]. Tiap lipatan data akan dilakukan pengukuran performa metode.

E. *Gaussian Naive Bayes*

Gaussian Naive Bayes adalah salah satu varian *Naive Bayes*. Metode ini bekerja pada tipe data angka dengan cara sebagai berikut [3]:

Menentukan nilai *mean* setiap atribut atau fitur terhadap kelas objek, dengan persamaan:

$$mean(x) = \frac{1}{n}\ x\ \sum_{i=1}^{n} x_i \quad (5)$$

Dimana, n merupakan jumlah data dari kelas objek pada atribut. Selanjutnya dihitung standar deviasi tiap kelas objek terhadap setiap atribut menggunakan persamaan:

$$StandardDeviation(x) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - mean(x))^2}{n-1}} \quad (6)$$

Dimana, x merupakan kelas objek tertentu dan n merupakan jumlah data pada kelas objek tersebut. Selanjutnya adalah menghitung *Prior Probability* untuk mengetahui probabilitas tiap kelas objek terhadap data seluruhnya menggunakan persamaan :

$$Prior\ probability\ C1 = \frac{\sum A \mid X}{\sum X} \quad (7)$$

Dimana, A merupakan kelas objek dan X merupakan atribut. Ketiga persamaan ini digunakan dalam pemodelan, sedangkan dalam pengujian akan digunakan persamaan Gaussian Probability Density Function (PDF) :

$$pdf(x, meansd) = \frac{1}{\sqrt{2\pi}sd}\ x\ e^{-(\frac{(x-mean)^2}{2sd^2})} \quad (8)$$

Dimana, persamaan ini akan menggunakan nilai *mean* dan standar deviasi yang diperoleh tiap atribut terhadap kelas objek. Jika kelas objek memiliki tiga kelas, maka nilai *mean* dan standar deviasi akan berjumlah tiga tiap atribut.

F. *K Nearest Neighbor*

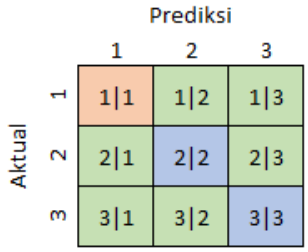
Algoritma ini bertujuan untuk mengklasifikasikan objek baru sesuai dengan atribut dan sampel pelatihan. Pada metode ini klasifikasi tidak memerlukan data latih untuk membentuk pola, itu hanya tergantung pada memori [8].

$$Euclidean\ Distance = \sqrt{\sum_{k=1}^{n}(X_k - Y_k)^2} \quad (9)$$

Dimana, X merupakan data latih sedangkan Y merupakan data uji yang akan diklasifikasi.

G. *Confusion Matrix*

Confusion Matrix adalah teknik untuk mengevaluasi hasil pengujian model. Teknik yang digunakan adalah dengan cara menggunakan tabel matriks. Data akan diklasifikasikan menjadi dua kelas, yang satu dianggap dianggap negatif [9



Gambar 1. *Confusion Matrix* 3 kelas
Sumber: [10]

3. METODE PENELITIAN

Penelitian ini disusun dalam beberapa tahapan. Berikut tahapan-tahapannya:

A. Pengumpulan Data

Pengumpulan data dilakukan dengan cara meminta data siswa pada SMA Muhammdiyah 3 jember. data yang digunakan pada penelitian ini berjumlah 320 data siswa.

B. *Preprocessing Data*

Pada tahap ini dilakukan dua langkah *preprocessing*, pertama pada permasalahan *Missing value* menggunakan metode median

Jurnal Smart Teknologi
Vol. 3, No. 3, Maret 2022, Halaman 261 – 266
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

subtution dan pada tahap normalisasi data menggunakan *Min-Max Normalization*.

C. Partisi Data

Pada data hasil *preprocessing* akan dibagi menjadi dua data yaitu, data latih dengan persentase 80% dan data uji dengan persentase 20%. 80% data latih akan dilakukan skenario pemodelan menggunakan metode *K Fold Cross Validation* dengan nilai k = 2, 4, 5, 8 dan 10.

D. Implementasi *GNB* & *KNN*

Impelementasi *Gaussian Naive Bayes* dan *K Nearest Neighbor* akan dilakukan pada pemodelan menggunakan skenario *K Fold Cross Validation*, penggunaan teknik *SMOTE (Synthetic Minority Oversampling Technique)* pada model dan pengujian data validasi.

E. Pengukuran

Pengukuran pada kedua metode tersebut adalah pengukuran tingkat akurasi, presisi dan *recall*.

F. Hasil

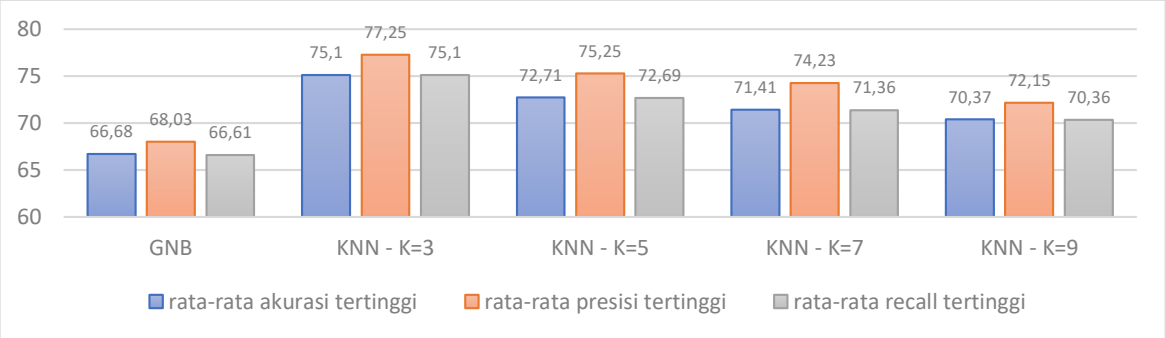
Hasil dari implementasi kedua metode akan dibandingkan. Hasil pada pemodelan dan pada pengujian data validasi. Analisis dilakukan pada hasil pemodelan terhadap hasil pengujian untuk mengukur fluktuatif performa metode.

4. HASIL PENELITIAN DAN PEMBAHASAN

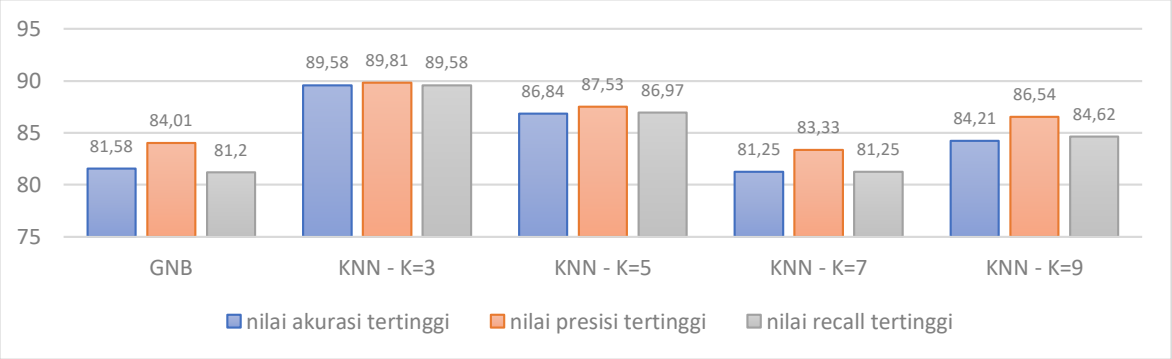
Dari partisi data, diperoleh perbandingan jumlah data uji 64 dan jumlah data latih 256. Kondisi data latih memiliki ketidakseimbangan dengan perbandingan jumlah kelas bahasa 127 siswa, kelas ipa 66 siswa dan kelas ips 63 siswa. Penerapan *SMOTE* mengakibatkan kondisi data menjadi seimbang dengan kondisi tiap kelas menjadi 127 data sama besar.

A. Pemodelan

Dari hasil pemodelan serta skenario pengujian menggunakan *K Fold Cross Validation* dengan nilai k = 2, 4, 5, 8 dan 10 diperoleh hasil pada gambar 2 di bawah ini.



Gambar 2. Hasil pengukuran rata-rata akurasi, presisi dan *recall* pada skenario *K Fold Cross Validation*
Sumber: Hasil perhitungan



Jurnal Smart Teknologi
Vol. 3, No. 3, Maret 2022, Halaman 261 – 266
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

Gambar 3. Hasil pengukuran akurasi, presisi dan *recall* tertinggi pada tiap *k Fold*
Sumber: Hasil perhitungan

Pada gambar 2 di atas, dapat dijelaskan bahwa pemodelan *K Nearest Neighbor* lebih baik dari pada pemodelan *Gaussian Naive Bayes*. Tingkat rata-rata tertinggi dari akurasi *K Nearest Neighbor* mencapai 75,1% yang diperoleh pada *Nearest Neighbor k* = 3. Tingkat rata-rata presisi tertinggi juga dimiliki *K Nearest Neighbor* dengan nilai 77,25% yang terletak pada *Nearest Neighbor k* = 3 dan rata-rata *recall* tertinggi juga dimiliki oleh *K Nearest Neighbor* dengan nilai 75,1% yang terletak pada *Nearest Neighbor k* = 3.

Dapat diketahui juga bahwa, hasil pemodelan *Gaussian Naive Bayes* lebih rendah dari *K Nearest Neighbor*. Nilai akurasi yang diperoleh metode *Gaussian Naive Bayes* sebesar 66, 68%, nilai presisi sebesar 68,3% dan nilai *recall* sebesar 66,61%. Dalam pemodelan ini, metode *K Nearest Neighbor* memiliki keunggulan dari pada metode *Gaussian Naive Bayes*.

B. Pengujian

Pengujian validasi dilakukan pada 64 data atau 20% data partisi dari data awal. Model yang digunakan adalah *Gaussian Naive Bayes* dan *K Nearest Neighbor* dengan nilai k = 3. Hal ini dipilih karena k = 3 merupakan nilai tetangga terdekat yang memiliki performa tertinggi dibandingkan *Nearest Neighbor* lainnya.

Tabel 1. Hasil pengukuran tingkat akurasi, presisi dan *recall* pada pengujian data validasi.

Metode	Akurasi	Presisi	*Recall*
GNB	**64.06**	**60.4**	**63.02**
***KNN* - K=3**	**64.06**	51.35	52.86

Sumber: Hasil perhitungan

Berdasarkan pada gambar 3 di atas, diketahui bahwa hasil pengukuran tingkat akurasi, presisi dan *recall* metode *Gaussian Naive Bayes* lebih baik dari pada metode *K Nearest Neighbor*. Pada pengukuran tingkat akurasi, metode *Gaussian Naive Bayes* memiliki performa yang sama yaitu 64%, pada pengukuran tingkat presisi *Gaussian Naive Bayes* unggul 9,05% dan pada pengukuran tingkat *recall Gaussian Naive Bayes* unggul 10,16%. Hal ini berbanding terbalik pada hasil pemodelan. Dimana, metode *K Nearest Neighbor* memiliki performa yang lebih baik.

5. PENUTUP

A. Kesimpulan

Berdasarkan rangkaian penelitian dan analisis yang dilakukan pada penelitian ini, diperoleh kesimpulan sebagai berikut:

1) Pemodelan pada metode *Gaussian Naive Bayes* memperoleh hasil sebagai beriku
   a. Rata-rata nilai akurasi tertinggi sebesar 66.68% dan nilai akurasi tertinggi dari semua *k Fold* sebesar 81.58%.
   b. Rata-rata nilai presisi tertinggi sebesar 68.03% dan nilai presisi tertinggi dari semua *k Fold* sebesar 84.01%.
   c. Rata-rata nilai *recall* tertinggi sebesar 66.61% dan nilai *recall* tertinggi dari semua *k Fold* sebesar 81.2%.
2) Pemodelan pada metode *K Nearest Neighbor* memperoleh hasil sebagai berikut
   a. Rata-rata nilai akurasi tertinggi sebesar 75.1% dan nilai akurasi tertinggi dari semua *k Fold* sebesar 89.58%.
   b. Rata-rata nilai presisi tertinggi sebesar 77.25% dan nilai presisi tertinggi dari semua *k Fold* sebesar 89.81%.
   c. Rata-rata nilai *recall* tertinggi sebesar 75.1% dan nilai *recall* tertinggi dari semua *k Fold* sebesar 89.58%.
3) Pengujian data uji validasi dari metode *Gaussian Naive Bayes* dan *K Nearest Neighbor* memperoleh hasil sebagai berikut
   a. Metode *Gaussian Naive Bayes* memperoleh nilai akurasi sebesar 64.06%, nilai presisi sebesar 60.4% dan nilai *recall* sebesar 63.02%.
   b. Metode *K Nearest Neighbor* memperoleh nilai akurasi sebesar 64.06%, nilai presisi sebesar 51.35% dan nilai *recall* sebesar 52.86%.
4) Dalam kasus ini dapat disimpulkan bahwa metode *K Nearest Neighbor* dalam

melakukan pemodelan klasifikasi penjurusan siswa SMA Muhammadiyah 3 Jember lebih baik dari metode *Gaussian Naive Bayes*, tetapi metode *Gaussian Naive Bayes* memiliki nilai performa yang lebih baik dalam pengujian data validasi yang dilakukan.

B. Saran

Penulis menyadari bahwa penelitian ini masih jauh dari kesempurnaan. Untuk itu penulis sangat terbuka jika pihak pembaca ingin mengembangkan penelitian ini ke depannya agar jauh lebih baik. Berikut poin-poin yang dapat dijadikan acuan untuk pengembangan ke depannya :

1. Pengembang dapat menambahkan data siswa lebih banyak agar terbentuk model yang lebih baik.
2. Pengembang dapat menggunakan metode *preprocessing* lain guna memperoleh hasil yang lebih baik.
3. Pengembang juga dapat menggunakan metode lain atau metode *ensemble* guna meningkatkan hasil klasifikasi pada siswa.

6. Referensi

A. Buku

[2] Turban, E., Rainer, R. K., & Potter, R. E. (2005). *Introduction to information technology* (Vol. 2, pp. 51-62). John Wiley & Sons.

[3] Jason Brownlee. (2016). Master *Machine Learning* Algorithms : Discover How They Work and Implement Them From Scratch.

[8] Kusrini, E. T. L. (2009). Algoritma *Data Mining*. Yogyakarta: Andi Offset.

[9] Bramer, M. (2007). *Measuring the Performance of a Classifier. Principles of Data Mining*.

B. Artikel Jurnal

[4] Eric R. Buhi, MPH, PhD, Patricia Goodson, PhD, Torsten B. Neilands, P. (2008). *Out of Sight, Not Out of Mind: Strategies for Handling Missing Data. Handling Missing Data*, 1(Handl. Missing Data), 83–92.

[5] Virmani, D., Taneja, S., & Malhotra, G. (2015). *Normalization based K means Clustering Algorithm*. arXiv preprint arXiv:1503.00900.

[6] Cost, S., & Salzberg, S. (1993). *A weighted Nearest Neighbor algorithm for learning with symbolic features. Machine Learning*, 10(1), 57-78.

[10] Ali, M., Son, D. H., Kang, S. H., & Nam, S. R. (2017). *An accurate CT saturation classification using a deep learning approach based on unsupervised feature extraction and supervised fine-tuning strategy*. Energies, 10(11), 1830.

C. Tesis atau Disertasi

[7] Haltuf, M. (2014). *Support Vector Machines for Credit Scoring. Thesis, Faculty of Finance University of Economics in Prague. Prague*.

D. Sumber Rujukan Dari *Website*

[1] SMA Muhammadiyah 3 Jember. (2021). Profil SMA Muhammadiyah 3 Jember. Pada laman https://smamuh3jbr.sch.id/tentang-sekolah/ diakses pada tanggal 1 July 2021.