ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 4, Nomor 2, Tahun 2015, Halaman 295 - 304
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

PERBANDINGAN DISKRIMINAN KUADRATIK KLASIK DAN DISKRIMINAN KUADRATIK ROBUST PADA KASUS PENGKLASIFIKASIAN PEMINATAN PESERTA DIDIK (Studi Kasus di SMA Negeri 1 Kendal Tahun Ajaran 2014/2015)

Laili Isna Nur Khiqmah[1], Moch. Abdul Mukid[2], Alan Prahutama[3]
[1]Mahasiswa Jurusan Statistika FSM UNDIP
[2,3]Staff Pengajar Jurusan Statistika FSM UNDIP

ABSTRAK

Diskriminan merupakan teknik statistik multivariat yang dapat digunakan untuk melakukan klasifikasi terhadap suatu observasi baru ke dalam suatu kelompok tertentu. Analisis diskriminan kuadratik terikat suatu asumsi observasi berdistribusi normal multiariat dan ketidaksamaan matriks varian kovarian. Analisis diskriminan kuadratik robust digunakan apabila observasi mengandung pencilan. Pengklasifikasian menggunakan analisis diskriminan kuadratik robust dengan penduga *Minimum Covariance Determinant* (MCD) pada data peminatan peserta didik SMA Negeri 1 Kendal yang mengandung pencilan memberikan hasil ketepatan klasifikasi sebesar 95,06% dengan persentase kesalahan klasifikasi sebesar 4,94% sedangkan analisis diskriminan kuadratik klasik menghasilkan ketepatan klasifikasi sebesar 92,59% dengan persentase kesalahan klasifikasi sebesar 7,41%. Dengan demikian analisis diskriminan kuadratik robust dengan penduga MCD lebih tepat digunakan pada kasus data yang mengandung pencilan.

Kata kunci : diskriminan, pencilan, robust, penduga MCD, klasifikasi.

ABSTRACT

Discriminant is a multivariate statistical technique that can be used to perform the classification new observation into a particular group. Quadratic discriminant analysis tied to an assumption of normal multivariate distributed observations and variance covariance matrix inequality. Robust quadratic discriminant analysis can be used if the observations contain outliers. Classification using robust quadratic discriminant analysis with the *Minimum Covariance Determinant* (MCD) estimator in the data specialization students of SMA Negeri 1 Kendal that containing outliers gives the results of the classification accuracy of 95,06% with a percentage of 4,94% classification error while generating the classical quadratic discriminant analysis classification accuracy of 92,59% with a percentage of 7,41% classification error. Thus a robust quadratic discriminant analysis with the MCD estimator is more appropriate in the case of the data which contains outliers.

Keywords : discriminant, outliers, robust, MCD estimators, classification.

1. PENDAHULUAN

1.1 Latar Belakang

Peminatan peserta didik pada jenjang Sekolah Menengah Atas (SMA) merupakan sesuatu yang dirasa penting dan perlu untuk diterapkan oleh pihak sekolah dengan berdasarkan Peraturan Menteri Pendidikan dan Kebudayaan RI Nomor 64 Tahun 2014 tentang Peminatan pada Pendidikan Menengah. Hal ini dikarenakan peminatan pada jenjang SMA merupakan suatu sarana untuk memberikan kesempatan pada peserta didik dalam mengembangkan kompetensi sikap, kompetensi pengetahuan, dan kompetensi keterampilan peserta didik sesuai dengan minat, bakat dan/atau kemampuan akademik dalam sekelompok mata pelajaran keilmuan.

SMA Negeri 1 Kendal merupakan salah satu Sekolah Menengah Atas (SMA) yang bertaraf RSBI (Rintisan Sekolah Bertaraf Internasional) di Kabupaten Kendal. Pihak sekolah SMA Negeri 1 Kendal telah menerapkan peminatan untuk para peserta didiknya sesuai dengan Peraturan Menteri Pendidikan dan Kebudayaan RI Nomor 64 Tahun 2014. Selain berdasarkan pada Peraturan Menteri Pendidikan dan Kebudayaan RI Nomor 64 Tahun 2014, pihak sekolah SMA Negeri 1 Kendal juga menyelenggarakan tes seleksi masuk dan tes psikotes sebagai bahan pertimbangan untuk mengklasifikasikan peserta didiknya berdasarkan peminatan yang ada.

Mengingat kesesuaian pengklasifikasian peserta didik pada kelompok peminatan merupakan hal yang penting untuk masa depan peserta didik tersebut, maka hal ini perlu untuk diperhatikan. Oleh karena itu diperlukan suatu metode statistik yang dapat digunakan untuk mengukur seberapa besar keakuratan pengelompokkan peminatan peserta didik SMA Negeri 1 Kendal. Salah satu metode statistik yang dapat digunakan untuk mengukur seberapa besar keakuratan pengelompokkan peminatan peserta didik SMA Negeri 1 Kendal yaitu analisis diskriminan.

Analisis diskriminan merupakan teknik multivariat yang berkaitan dengan pemisahan pengamatan dalam kelompok yang berbeda dan mengalokasikan pengamatan ke dalam suatu kelompok yang telah ditetapkan sebelumnya (Johnson dan Wichern, 2007). Dengan analisis diskriminan diharapkan dapat diketahui seberapa besar kesalahan klasifikasi yang dihasilkan dalam mengelompokkan suatu pengamatan pada kelompok tertentu. Salah satu analisis diskriminan yaitu analisis diskriminan kuadratik. Pada analisis diskriminan kuadratik terikat suatu asumsi bahwa data pengamatan harus memenuhi asumsi normal multivariat dan mempunyai matriks varians kovarians yang berbeda antar kelompoknya. Apabila pada data pengamatan mengandung pencilan maka dapat digunakan metode robust yaitu analisis diskriminan kuadratik robust.

Berdasarkan latar belakang di atas, maka penulis ingin mengkaji hasil kesalahan klasifikasi menggunakan analisis diskriminan kuadratik klasik dan analisis diskriminan kuadratik robust pada data peminatan peserta didik SMA Negeri 1 Kendal pada tahun ajaran 2014/2015 ke dalam kelompok peminatan yang ada. Hal itulah yang mendasari penulisan laporan tugas akhir yang berjudul "Perbandingan Diskriminan Kuadratik Klasik dan Diskriminan Kuadratik Robust pada Kasus Pengklasifikasian Peminatan Peserta Didik (Studi Kasus di SMA Negeri 1 Kendal Tahun Ajaran 2014/2015)".

1.2 Tujuan

Tujuan penelitian dari penulisan Tugas Akhir ini adalah :

1. Menentukan skor kuadratik klasik tiap pengamatan di setiap kelompok peminatan peserta didik SMA Negeri 1 Kendal.
2. Menentukan skor kuadratik robust tiap pengamatan di setiap kelompok peminatan peserta didik SMA Negeri 1 Kendal.
3. Menentukan besar kesalahan hasil pengklasifikasian kelompok peminatan peserta didik SMA Negeri 1 Kendal menggunakan analisis diskriminan kuadratik klasik dan analisis diskriminan kuadratik robust.
4. Membandingkan kesalahan klasifikasi kedua metode menggunakan kriteria nilai APER.

2. TINJAUAN PUSTAKA

2.1 Penentuan Kelompok Peminatan

Peminatan adalah program kurikuler yang disediakan untuk mengakomodasi pilihan minat, bakat dan/atau kemampuan peserta didik dengan orientasi pemusatan, perluasan, dan/atau pendalaman mata pelajaran dan/atau muatan kejuruan (Peraturan Menteri Pendidikan dan Kebudayaan RI Nomor 64 Tahun 2014).

Terdapat beberapa indikator yang digunakan untuk melakukan pengelompokkan diantaranya meliputi rata-rata nilai mata pelajaran Bahasa Indonesia, rata-rata nilai mata pelajaran Bahasa Inggris, rata-rata nilai mata pelajaran matematika, rata-rata nilai mata pelajaran IPA, rata-

rata nilai mata pelajaran IPS, dan nilai tes psikotes. Nilai rata-rata setiap mata pelajaran diambil dari nilai rapor SMP/MTs atau yang sederajat, nilai Ujian Nasional SMP/MTs dan nilai tes seleksi masuk.

2.2 Analisis Diskriminan Kuadratik Klasik

Fungsi normal multivariat dapat dibentuk melalui fungsi densitas probabilitas untuk distribusi normal univariat dengan rata-rata μ dan variansi σ yang secara umum dapat ditulis dengan notasi $N(\mu, \sigma)$:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2; -\infty \leq x \leq \infty \quad (1)$$

untuk kasus normal multivariat, maka fungsi densitas probabilitas untuk distribusi normal univariat dengan rata-rata μ dan variansi σ diperluas menjadi :

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)^T(\sigma^2)^{-1}(x-\mu) \quad (2)$$

Sehingga untuk vektor $\boldsymbol{x}$ dengan d buah variabel bebas dapat ditulis menjadi :

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \quad (3)$$

dimana $\boldsymbol{\mu} = E(X)$ dan $\boldsymbol{\Sigma} = Var\ Kov(X)$.

Dengan menambahkan konstanta $(2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}$, maka fungsi densitas normal multivariat d dimensi untuk vektor random $\boldsymbol{X} = [X_1, X_2, .., X_d]^T$ mempunyai bentuk umum

$$f(x) = (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}\,)\right); -\infty \leq x_d \leq \infty \quad (4)$$

yang dapat dinotasikan dengan $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Menurut Johnson dan Wichern (2007), pada kasus multivariat, vektor peubah acak $\mathbf{X}'=[\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_d]$ mengikuti fungsi densitas probabilitas $f(\boldsymbol{x})$. Apabila data sampel berdistribusi normal multivariat, maka fungsi kepadatan peluang $(f_i(\boldsymbol{x}))$ untuk data berdistribusi normal multivariat dengan vektor rata-rata $\boldsymbol{\mu}_i$ dan matriks varian kovarian $\boldsymbol{\Sigma}_i$ dapat ditulis :

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}{2}\right) \quad (5)$$

Jika p_i adalah peluang awal dari kelompok π_i dengan $i = 1,2, \dots, g$ dan $C(k|i)$ besarnya nilai kesalahan mengelompokkan suatu observasi anggota kelompok π_k, yang berasal dari π_i untuk $k, i = 1,2, \dots, g$. Maka diperoleh suatu aturan pengelompokkan observasi dalam kelompok π_k terhadap kelompok π_i:

$$p_k f_k(\boldsymbol{x}) > p_i f_i(\boldsymbol{x}) \text{ dengan } k \neq i \quad (6)$$

$$\ln p_k f_k(\boldsymbol{x}) > \ln p_i f_i(\boldsymbol{x}) \quad (7)$$

Untuk selanjutnya dicari $\ln p_k f_k(\boldsymbol{x})$ dengan memasukkan persamaan 5 diperoleh :

$$\Leftrightarrow \ln p_k f_k(\boldsymbol{x}) = \ln p_k \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}}\ exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}{2}\right) \quad (8)$$

$$\Leftrightarrow \ln p_k f_k(\boldsymbol{x}) = \ln p_k - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_k|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) \quad (9)$$

Karena konstanta $\frac{d}{2}\ln(2\pi)$ bernilai sama untuk semua kelompok maka dapat diabaikan, sehingga persamaan (13) menjadi :

$$\Leftrightarrow \ln p_k f_k(\boldsymbol{x}) = \ln p_k - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_k|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) \quad (10)$$

$$\Leftrightarrow \ln p_k f_k(\boldsymbol{x}) = max_i \ln p_i f_i(\boldsymbol{x}) \quad (11)$$

Berdasarkan pengelompokkan pada distribusi normal multivariat, maka skor diskriminan kuadratik untuk g kelompok didefinisikan sebagai :

$$d_i^Q(\boldsymbol{x}) = \ln p_i - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| \quad (12)$$

Menurut Johnson dan Wichern (2007), $\boldsymbol{\mu}_i$ dan $\boldsymbol{\Sigma}_i$ merupakan parameter yang tidak diketahui sehingga harus dicari estimasi dari sampelnya. Sehingga, estimasi dari skor diskriminan kuadratik menjadi:

$$d_i^Q(x) = \ln p_i - \frac{1}{2}(\boldsymbol{x} - \overline{\boldsymbol{x}}_i)^T \boldsymbol{S}_i^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}_i) - \frac{1}{2}\ln|\boldsymbol{S}_i| \quad (13)$$

sehingga taksiran pengelompokkannya adalah dengan menempatkan $\boldsymbol{x}$ ke dalam kelompok π_k jika skor diskriminan kuadratik :

$$d_k^Q(\boldsymbol{x}) = \text{maks } \{d_1^Q(\boldsymbol{x}), d_2^Q(\boldsymbol{x}), \ldots, d_g^Q(\boldsymbol{x})\} \quad (14)$$

Dua pendekatan umum yang dapat digunakan untuk mengestimasi probabilitas anggota p_i, yaitu $p_i = \frac{1}{g}$ yang mana diasumsikan p_i sama untuk semua kelompok, dan $p_i = \frac{n_i}{n}$ untuk setiap kelompok dengan p_i diestimasi sebagai frekuensi relatif dari observasi pada setiap kelompok (Hubert dan Driessen, 2002).

2.3 Analisis Diskriminan Kuadratik Robust

2.3.1 Penduga *Minimum Covariance Determinant* (MCD)

Minimum Covariance Determinant (MCD) merupakan salah satu penduga robust untuk analisis diskriminan kuadratik yang cukup populer. Penduga MCD merupakan pasangan $\boldsymbol{t}_m$ dan $\boldsymbol{C}_m$ yang masing-masing merupakan vektor rata-rata sampel dan matriks kovarian dari data sebanyak h observasi. Rousseeuw dan Van Driessen (1999) memperkenalkan algoritma baru yang disebut *fast*-MCD yang mampu bekerja lebih cepat dan mampu menangani himpunan data yang sangat besar. *fast*-MCD merupakan algoritma baru yang didasarkan pada algoritma C-*Step* (Maronna *et al.*, 2006). Dimana algoritma Teorema C-*Step* adalah sebagai berikut (Rousseeuw dan Driessen, 1999) :

Dari sebanyak n observasi dan d variabel observasi $X_n = \{X_1, X_2, \ldots, X_n\}$. Tetapkan $\boldsymbol{H}_1 \subset \{1, 2, \ldots, n\}$ dengan $|\boldsymbol{H}_1| = h$ merupakan subsampel sebanyak h anggota himpunan, dimana $h = \left\lfloor \frac{n+d+1}{2} \right\rfloor$. Kemudian hitung $\boldsymbol{t}_1$ dan $\boldsymbol{C}_1$ yang masing-masing dirumuskan sebagai :

$$\boldsymbol{t}_1 = \frac{1}{h}\sum_{j \in H_1} \boldsymbol{x}_j \quad (15)$$

$$\boldsymbol{C}_1 = \frac{1}{h}\sum_{j \in H_1} (\boldsymbol{x}_j - \boldsymbol{t}_1)(\boldsymbol{x}_j - \boldsymbol{t}_1)^T \quad (16)$$

Apabila $\det(\boldsymbol{C}_1) \neq 0$, hitung jarak Mahalanobis sebagai jarak relatif dengan rumus :

$$d_{j1} = \sqrt{(\boldsymbol{x}_j - \boldsymbol{t}_1)^T \boldsymbol{C}_1^{-1}(\boldsymbol{x}_j - \boldsymbol{t}_1)} \quad (17)$$

dengan $j = 1,2, \ldots, n$. Kemudian tentukan anggota himpunan $\boldsymbol{H}_2$ yang memenuhi $\{d_1(j); j \in \boldsymbol{H}_2\} = \{(d_{j1})_{1:n}, \ldots, (d_{j1})_{h:n}\}$, dimana $(d_{j1})_{1:n} \leq (d_{j1})_{2:n} \leq \ldots \leq (d_{j1})_{n:n}$ merupakan urutan jarak dari sebanyak n observasi. Berdasarkan anggota himpunan $\boldsymbol{H}_2$ hitung $\boldsymbol{t}_2$ dan $\boldsymbol{C}_2$ menggunakan rumus seperti pada persamaan (15) dan (16) maka :

$$\det(\boldsymbol{C}_2) \leq \det(\boldsymbol{C}_1)$$

Iterasi dilakukan sampai ditemukan himpunan $\boldsymbol{H}_k$ yang konvergen dan memiliki determinan matriks kovarian terkecil. Dimana $\det(\boldsymbol{C}_2)$ akan sama dengan $\det(\boldsymbol{C}_1)$ jika dan hanya jika $\boldsymbol{t}_2 = \boldsymbol{t}_1$ dan $\boldsymbol{C}_2 = \boldsymbol{C}_1$.

Prosedur untuk menghitung penduga MCD dengan algoritma *fast*-MCD dilakukan dengan langkah-langkah sebagai berikut (Rousseeuw dan Driessen, 1999) :

1. Dari sebanyak *n* observasi diambil sejumlah *h* observasi yang memenuhi $\left\lfloor \frac{n+d+1}{2} \right\rfloor \leq h \leq n$. Ditetapkan nilai standar *h* yaitu $h = \left\lfloor \frac{n+d+1}{2} \right\rfloor$.
2. Jika $h = n$, maka penduga $\boldsymbol{t}_{MCD}$ dan $\boldsymbol{C}_{MCD}$ masing-masing merupakan rata-rata dan matriks kovarian dari seluruh observasi.

3. Jika $h < n$, $d \geq 2$ dan n berukuran kecil (misal $n \leq 600$), maka :
   a. Definisikan himpunan pertama sebagai $\boldsymbol{H}_1$ dengan cara mengambil secara acak $(d+1)$ himpunan bagian K. Kemudian hitung $\boldsymbol{t}_0$ dan $\boldsymbol{C}_0$ yang masing-masing merupakan vektor rata-rata dan matrik kovarian dari K. [Jika det($\boldsymbol{C}_0$) $= 0$, maka perluas K dengan menambahkan observasi acak lainnya hingga diperoleh det($\boldsymbol{C}_0$) $>$ 0]. Selanjutnya hitung jarak $d_0^2 = \left(\boldsymbol{x_j} - \boldsymbol{t}_0\right)^T \boldsymbol{C}_0^{-1}\left(\boldsymbol{x_j} - \boldsymbol{t}_0\right)$ dimana $j = 1, \ldots, n$. Urutkan d_0^2 sedemikian rupa ke dalam $d_0(\pi(1)) \leq \ldots \leq d_0(\pi(n))$ dan tentukan sebagai $\boldsymbol{H}_1 = \{\pi(1), \ldots, \pi(h)\}$.
   b. Lakukan dua kali C-Step terhadap $\boldsymbol{H}_1$ sampai menghasilkan $\boldsymbol{H}_3$.
   c. Ulangi langkah a dan b sebanyak 500 kali.
   d. Pilih sepuluh hasil dari $\boldsymbol{H}_3$ dengan nilai det($\boldsymbol{C}_3$) terkecil.
   e. Lakukan C-Step sampai konvergen hingga diperoleh hasil $(\boldsymbol{t}_m, \boldsymbol{C}_m)$ dengan det($\boldsymbol{C}_m$) terkecil.
4. Berdasarkan nilai $\boldsymbol{t_m}$ dan $\boldsymbol{C_m}$ dari data H_m selanjutnya diberi bobot

$$w_j = \begin{cases} 1\,;\, jika\ (x_j - \boldsymbol{t_m})^T \boldsymbol{C}_m^{-1}\,(x_j - \boldsymbol{t_m}) \leq X^2_{p,(\alpha)} \\ \qquad\qquad 0\,;\ jika\ lainnya \end{cases} \tag{18}$$

5. Berdasarkan nilai perhitungan bobot pada persamaan (18), maka MCD dihitung sebagai :

$$\boldsymbol{t}_{MCD} = \frac{\sum_{j=1}^{n} w_j \boldsymbol{x}_j}{\sum_{j=1}^{n} w_j} \tag{19}$$

$$\boldsymbol{C}_{MCD} = \frac{\sum_{j=1}^{n} w_j (\boldsymbol{x}_j - \boldsymbol{t}_{MCD})(\boldsymbol{x}_j - \boldsymbol{t}_{MCD})^T}{\sum_{j=1}^{n} w_j - 1} \tag{20}$$

2.3.2 Pendeteksian Pencilan

Data pencilan adalah data yang memiliki karakterstik khusus yang terlihat sangat berbeda jauh dari observasi-observasi lainnya dan muncul dalam bentuk nilai ekstrim baik untuk sebuah variabel tunggal atau variabel kombinasi (Ghozali, 2005). Jarak robust merupakan uuran jarak yang dapat mendeteksi pencilan lebih baik dari pada jarak Mahalanobis (Rousseeuw dan Driessen, 1999). Jika dipunyai data multivariat d dimensi dengan sampel x_j $(j = 1, \ldots, n)$ maka jarak robust didefinisikan sebagai (Rousseeuw dan Van Zomeren, 1990):

$$DR_{ij}^2 = \left(\boldsymbol{x_{ij}} - \boldsymbol{t_{m(i)}}\right)^T \boldsymbol{C}_{m(i)}^{-1} \left(\boldsymbol{x_{ij}} - \boldsymbol{t_{m(i)}}\right) \tag{21}$$

Apabila observasi $\boldsymbol{x_{ij}}$ mempunyai nilai $DR_{ij}^2 > x^2_{\alpha;d}$ maka observasi $\boldsymbol{x_{ij}}$ merupakan pencilan.

2.3.3 Diskriminan Kuadratik Robust

Menurut Hubert dan Driessen (2002), skor diskriminan kuadratik dengan menggunakan penduga *robust* MCD diperoleh dengan mengganti penduga vektor rata-rata dan matrik kovariansi dengan penduga MCD. Adapun formulasinya sebagai berikut :

$$d_i^{QMCD}(\boldsymbol{x}) = \ln p_i - \tfrac{1}{2}\left(\boldsymbol{x} - \boldsymbol{t}_{MCD(i)}\right)^T \boldsymbol{C}_{MCD(i)}^{-1}\left(\boldsymbol{x} - \boldsymbol{t}_{MCD(i)}\right) - \tfrac{1}{2}\ln\left|\boldsymbol{C}_{MCD(i)}\right| \tag{22}$$

Observasi $\boldsymbol{x_{ij}}$ akan masuk ke dalam kelompok π_k jika skor diskriminan kuadratik

$$d_k^{QMCD}(\boldsymbol{x}) = \text{maks}\left\{d_1^{QMCD}(\boldsymbol{x}), d_2^{QMCD}(\boldsymbol{x}), \ldots, d_g^{QMCD}(\boldsymbol{x})\right\} \tag{23}$$

2.4 Masalah Klasifikasi (*Misclassification)*

Misclassification merupakan kesalahan dari pengklasifikasian suatu observasi baru ke dalam suatu kelompok (Johnson dan Wichern, 2007). Salah satu metode yang dapat digunakan untuk menghitung probabilitas kesalahan klasifikasi adalah APER (Johnson dan Wichern, 2007).

Tabel 2. Matriks Konfusi Hasil Klasifikasi

		Kelompok hasil Klasifikasi			
		π_1	π_2	...	π_g
Kelompok asli	π_1	n_{11}	n_{12}	...	n_{1g}
	π_2	n_{21}	n_{22}	...	n_{2g}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	π_g	n_{g1}	n_{g2}	...	n_{gg}

dengan :

$n_1 + n_2 + \ldots + n_g = n$

Maka statistik uji nilai $APER$ dirumuskan :

$$APER = \left(\frac{\sum_i \sum_j n_{ij}}{n}\right) x\ 100\% \text{ dengan } i \neq j \tag{24}$$

3. METODOLOGI PENELITIAN

3.1 Jenis dan Sumber Data

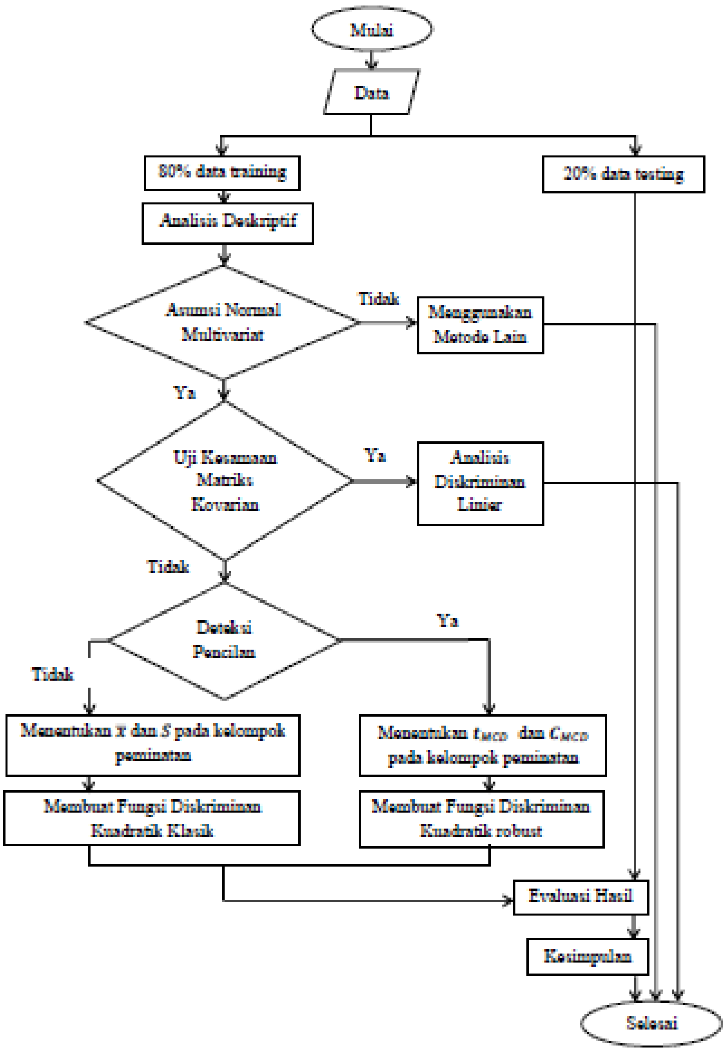
Data yang digunakan dalam penelitian tugas akhir ini adalah data sekunder yang bersumber dari SMA Negeri 1 Kendal berupa data nilai rapor SMP/MTs atau yang sederajat, nilai Ujian Nasional SMP/MTs dan nilai hasil tes seleksi peserta didik SMA Negeri 1 Kendal tahun ajaran 2014/2015 yang meliputi mata pelajaran Bahasa Indonesia, Bahasa Inggris, Matematika, IPA, IPS, dan nilai tes psikotes serta data peminatan peserta didik sebanyak 401 data, dengan rincian yaitu 200 observasi termasuk jurusan Matematika dan IPA, 138 observasi termasuk jurusan IPS, dan 63 observasi termasuk jurusan Bahasa dan Budaya.

3.2 Variabel Penelitian

Dalam analisis data digunakan sebanyak enam variabel yaitu Rata-rata nilai Bahasa Indonesia, Rata-rata nilai Bahasa Inggris, Rata-rata nilai Matematika, Rata-rata nilai IPA, Rata-rata nilai IPS, dan Nilai tes psikotes peserta didik, serta satu variabel terikat berupa data peminatan peserta didik SMA Negeri 1 Kendal tahun pelajaran 2014/2015.

3.4 Tahapan Analisis Data

Berikut merupakan diagram alir analisis sesuai dengan tahapan analisis yang telah dipaparkan :



Gambar 2. Diagram Alir Analisis

4. HASIL DAN PEMBAHASAN

4.1. Statistik Deskriptif

Berdasarkan data peminatan peserta didik SMA Negeri 1 Kendal tahun ajaran 2014/2015, sebesar 50% siswa masuk pada kelompok IPA, sebesar 34% siswa masuk pada kelompok IPS dan sisanya sebesar 16% siswa masuk pada kelompok bahasa. peserta didik yang masuk pada kelompok peminatan Ilmu Pengetahuan Alam (IPA) cenderung mempunyai nilai rata-rata baik untuk mata pelajaran Bahasa Indonesia, Bahasa Inggris, Matematika, IPA, IPS maupun nilai tes psikotes lebih tinggi dari pada Peserta didik yang masuk pada kelompok peminatan Ilmu Pengetahuan Sosial (IPS) maupun kelompok peminatan bahasa.

4.2. Uji Asumsi Distribusi Normal Multivariat

Pengujian asumsi distribusi normal multivariat dilakukan pada setiap kelompok peminatan. Pada kelompok peminatan IPA diperoleh hasil bahwa nilai p-value (0,2962) > α (0,05) dan secara visual membentuk pola grafik yang cenderung mengikuti garis lurus sehingga dapat dikatakan bahwa data peminatan kelompok IPA berdistribusi normal multivariat. Pada kelompok peminatan IPS diperoleh hasil bahwa nilai p-value (0,6012) > α (0,05), Secara visual membentuk pola grafik yang cenderung mengikuti garis lurus sehingga dapat dikatakan bahwa data peminatan kelompok IPS berdistribusi normal multivariat. Pada kelompok peminatan bahasa diperoleh hasil bahwa nilai p-value (0,6064) > α (0,05), Secara visual membentuk pola grafik yang cenderung mengikuti garis lurus sehingga dapat dikatakan bahwa data peminatan kelompok bahasa berdistribusi normal multivariat.

4.3. Uji Kesamaan Matriks Varian Kovarian

Pada pengujian asumsi kesamaan matriks varian kovarian diperoleh hasil bahwa nilai sig (0,000) < α (0,05), sehingga dapat diartikan bahwa terdapat paling sedikit dua matriks varian kovarian berbeda antar kelompok.

4.6. Analisis Diskriminan Kuadratik Klasik

Pada analisis diskriminan kuadratik klasik, langkah pertama yang harus dilakukan adalah menentukan vektor rata-rata dan matriks varian kovarian tiap kelompok peminatan. Setelah mendapatkan vektor rata-rata dan matriks varian kovarian, kemudian membentuk fungsi diskriminan kuadratik klasik pada masing-masing kelompok peminatan dalam bentuk sebagai berikut:

$$d_{IPA}^{Q}(\boldsymbol{x}) = \ln 0{,}5 - \frac{1}{2}(\boldsymbol{x} - \overline{\boldsymbol{x}}_{IPA})^{T}\boldsymbol{S}_{IPA}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}_{IPA}) - \frac{1}{2}\ln 297236498$$

$$d_{IPS}^{Q}(\boldsymbol{x}) = \ln 0{,}3438 - \frac{1}{2}(\boldsymbol{x} - \overline{\boldsymbol{x}}_{IPS})^{T}\boldsymbol{S}_{IPS}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}_{IPS}) - \frac{1}{2}\ln 134411289$$

$$d_{bahasa}^{Q}(\boldsymbol{x}) = \ln 0{,}1563 - \frac{1}{2}(\boldsymbol{x} - \overline{\boldsymbol{x}}_{bahasa})^{T}\boldsymbol{S}_{bahasa}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}_{bahasa}) - \frac{1}{2}\ln 125364658$$

Fungsi diskriminan kuadratik digunakan untuk menghitung skor kuadratik setiap observasi di setiap kelompok peminatan. Setelah diperoleh skor kuadratik, masing – masing observasi diklasifikasikan ke dalam kelompok peminatan IPA, IPS atau bahasa. Pengklasifikasian peserta didik SMA Neger 1 Kendal tahun ajaran 2014/2015 menggunakan fungsi diskriminan kuadratik klasik memberikan hasil Pada kelompok IPA terdapat sebanyak 2 observasi salah diklasifikasikan ke kelompok IPS, pada kelompok IPS terdapat sebanyak 3 observasi salah diklasifikasikan ke kelompok IPA, dan pada kelompok bahasa terdapat sebanyak 1 observasi salah diklasifikasikan ke kelompok IPS.

4.4. Pendeteksian Pencilan

Adanya Pencilan data akan mengakibatkan hasil pengelompokkan menggunakan diskriminan kuadartik klasik tidak optimal. Apabila taraf signifikansi yang digunakan adalah 5%, maka terdapat data pencilan sebanyak 19 observasi (11,875%) pada kelompok IPA, sebanyak 11 observasi (10%) pada kelompok IPS dan sebanyak 13 (26%) pada kelompok bahasa. Observasi yang merupakan pencilan selanjutnya diberi bobot 0 sedangkan observasi yang bukan merupakan pencilan diberi bobot 1. Pada kelompok peminatan IPA, penduga MCD dihasilkan dari sebanyak 141 pengamatan yang bukan merupakan pencilan. Pada kelompok peminatan IPS, penduga MCD dihasilkan dari sebanyak 99 pengamatan yang bukan merupakan pencilan. Sementara itu, pada kelompok peminatan bahasa, penduga MCD dihasilkan dari sebanyak 37 pengamatan yang bukan merupakan pencilan.

4.7. Analisis Diskriminan Kuadratik *Robust*

Pada analisis diskriminan kuadratik *robust*, langkah pertama yang harus dilakukan sama seperti pada analisis diskriminan kuadratik klasik yaitu menentukan vektor rata-rata dan matriks varian kovarian tiap kelompok peminatan. Setelah mendapatkan vektor rata-rata dan matriks varian kovarian, kemudian membentuk fungsi diskriminan kuadratik *robust* pada masing-masing kelompok peminatan dalam bentuk sebagai berikut:

$$d_{IPA}^{QMCD}(\boldsymbol{x}) = \ln 0{,}5 - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{t}_{MCD(IPA)})^{T}\boldsymbol{C}_{MCD(IPA)}^{-1}(\boldsymbol{x} - \boldsymbol{t}_{MCD(IPA)}) - \frac{1}{2}\ln 94980856$$

$$d_{IPS}^{QMCD}(\boldsymbol{x}) = \ln 0{,}3438 - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{t}_{MCD(IPS)})^T \boldsymbol{C}_{MCD(IPS)}^{-1}(\boldsymbol{x} - \boldsymbol{t}_{MCD(IPS)}) - \frac{1}{2}\ln 69819802$$
$$d_{bahasa}^{QMCD}(\boldsymbol{x}) = \ln 0{,}1563 - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{t}_{MCD(bahasa)})^T \boldsymbol{C}_{MCD(bahasa)}^{-1}(\boldsymbol{x} - \boldsymbol{t}_{MCD(bahasa)}) - \frac{1}{2}\ln 16888116$$

Fungsi diskriminan kuadratik digunakan untuk menghitung skor kuadratik setiap observasi di setiap kelompok peminatan. Setelah diperoleh skor kuadratik, masing – masing observasi diklasifikasikan ke dalam kelompok peminatan IPA, IPS atau bahasa. Pengklasifikasian peserta didik SMA Neger 1 Kendal tahun ajaran 2014/2015 menggunakan fungsi diskriminan kuadratik klasik memberikan hasil Pada kelompok IPA terdapat sebanyak 2 observasi salah diklasifikasikan ke kelompok IPS, pada kelompok IPS terdapat sebanyak 1 observasi salah diklasifikasikan ke kelompok IPA, dan pada kelompok bahasa terdapat sebanyak 1 observasi salah diklasifikasikan ke kelompok IPA.

4.8. Evaluasi Hasil Pemisahan

Perhitungan kesalahan klasifikasi seluruh pengamatan pada data testing untuk mengetahui seberapa besar kesalahan klasifikasi menggunakan analisis diskriminan kuadratik klasik dengan kriteria nilai APER sebagai berikut :

$$APER = \left(\frac{2+3+1}{81}\right) x\ 100\% = \left(\frac{6}{81}\right) x\ 100\% = 7{,}41\%$$

Sehingga dapat diketahui tingkat ketepatan klasifikasi data sebesar 92,59%.

Sementara itu, perhitungan kesalahan klasifikasi seluruh data pada data testing untuk mengetahui seberapa besar kesalahan klasifikasi yang dihasilkan menggunakan analisis diskriminan kuadratik *robust* dengan kriteria nilai APER sebagai berikut :

$$APER = \left(\frac{2+1+1}{81}\right) x\ 100\% = \left(\frac{4}{81}\right) x\ 100\% = 4{,}94\%$$

Sehingga dapat diketahui tingkat ketepatan klasifikasi data training yaitu sebesar 95,06%.

5. KESIMPULAN

Berdasarkan analisis dan pembahasan yang telah dipaparkan pada bab IV, dapat diperoleh beberapa kesimpulan sebagai berikut :

1. Pengklasifikasian menggunakan analisis diskriminan kuadratik klasik pada kelompok peminatan peserta didik SMA Negeri 1 Kendal tahun ajaran 2014/2015 menunjukkan bahwa sebanyak 38 peserta didik diklasifikasikan dengan benar ke dalam kelompok peminatan IPA, sebanyak 25 peserta didik diklasifikasikan dengan benar ke dalam kelompok peminatan IPS, dan sebanyak 12 peserta didik diklasifikasikan dengan benar ke dalam kelompok peminatan bahasa. Sehingga terdapat sebanyak 6 peserta didik salah diklasifikasikan.
2. Pengklasifikasian menggunakan analisis diskriminan kuadratik robust pada kelompok peminatan peserta didik SMA Negeri 1 Kendal tahun ajaran 2014/2015 menunjukkan bahwa sebanyak 38 peserta didik diklasifikasikan dengan benar ke dalam kelompok peminatan IPA, sebanyak 27 peserta didik diklasifikasikan dengan benar ke dalam kelompok peminatan IPS, dan sebanyak 12 peserta didik diklasifikasikan dengan benar ke dalam kelompok peminatan bahasa. Sehingga terdapat sebanyak 4 peserta didik salah diklasifikasikan.
3. Analisis diskriminan kuadratik robust dengan penduga MCD lebih tepat digunakan pada kasus pengklasifikasian peminatan peserta didik SMA Negeri 1 Kendal tahun ajaran 2014/2015 yang mengandung pencilan. Analisis diskriminan kuadratik robust memberikan hasil ketepatan klasifikasi yang lebih baik yaitu sebesar 95,06% dengan persentase kesalahan klasifikasi sebesar 4,94% dibandingkan dengan analisis diskriminan kuadratik klasik yang menghasilkan ketepatan klasifikasi lebih kecil yaitu sebesar 92,59% dengan persentase kesalahan klasifikasi sebesar 7,41%.

6. DAFTAR PUSTAKA

Daniel, W. W. 1989. *Statistika Non Parametrik Terapan.* Jakarta: PT Gramedia.

Ghozali, I. 2005. *Aplikasi Analisis Multivariate dengan Program IBM SPSS19.* Semarang: Badan Penerbit Universitas Diponegoro.

Hair, J. F., dkk. 2006. *Multivariate Data Analysis.* Sixth Edition. Pearson Education PrenticeHall, Inc.

Hubert, M. and Driessen, K.V. 2004. *Fast and robust discriminant analysis*. Computational Statistics & Data Analysis. Vol. 45. 301-320.

Johnson, R. A. dan Wichern, D. W. 2007. *Applied Multivariat Statistical Analysis,* 6^{th}*Edition.* New Jersey: Prentice Hall, Englewood Cliffs.

Maronna, R. A., Martin, R. D., dan Yohai, V. J. 2006. *Robust Statistics: Theory and Methods.* New York: John Wiley and Sons, Ltd.

Republik Indonesia. 2014. *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 64 Tahun 2014 tentang Peminatan pada Pendidikan Menengah*. Jakarta: Depdiknas.

Rousseeuw, P. J. and Driessen, K. V. 1999. *A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics.* Vol. 41, 212-223.

Rousseeuw, P. J. and Van Zomeren, B. C. 1990. *Unmasking Multivariate Outliers and Leverage Points.* Journal of American Statistical Association. Vol. 85, 633-639.

Seber, G.A.F. 2004. *Multivariat Observation.* New Jersey: John Wiley and Sons.

Suryana. 2008. *Perbandingan Kinerja Penaksir Kekar MCD dan MWCD dalam Analisis Diskriminan Kuadratik*. Tesis. Surabaya: Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Surabaya.

Werner, M. 2003. *Identification of Multivariate Outliers in Large Data Set*. Tesis. Denver: University of Colorado.

Widarjono, A. 2010. *Analisis Statistika Multivariat Terapan*. Yogyakarta: UPP STIM YKPN.