Analisis Kualitas Bahan Baku Tebu Melalui Teknik Pengklasteran dan Klasifikasi Kadar Gula Sebelum Giling (Studi Kasus Pabrik Gula PT. XYZ)

Merisa Khristanti Febriana Hanka dan Budi Santosa
Departemen Teknik Sistem dan Industri, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: budi_s@ie.its.ac.id

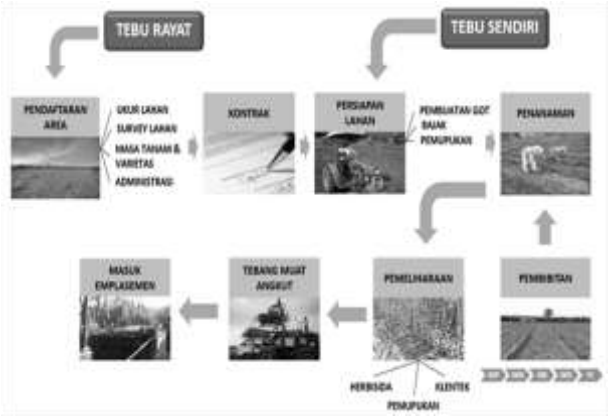
***Abstrak*—Sub sektor Perkebunan berkontribusi sebesar 3,27% terhadap Produk Domestik Bruto (PDB) sebagai urutan pertama di sektor pertanian pada tahun 2019. Tebu merupakan salah satu komoditi perkebunan yang mempunyai peran strategis dalam perekonomian di Indonesia. Salah satu Pabrik Gula (PG) milik PT. XYZ yang berlokasi di Jawa Timur memiliki mitra lebih dari 2000 petani tebu. Analisis mutu BBT PT. XYZ berdasarkan kriteria uji visual masih membuat adanya kemungkinan terjadi bias atau penyimpangan yang dilakukan terhadap analisis kadar gula. Menganalisis mutu BBT berdasarkan uji kadar gula diperlukan untuk mengetahui bagaimana klasifikasi dari kualitas BBT yang dikirim oleh mitra petani tebu sebelum masuk ke proses giling. Sistem ini dapat digunakan sebagai evaluasi kinerja petani tebu untuk dapat meningkatkan kualitas BBT serta keuntungan dalam bentuk bagi hasil antara petani tebu dan perusahaan. Penentuan mutu gula BBT bisa diatasi dengan pendekaan data mining yaitu teknik pengklasteran *Hierarchical K-Means Clustering* berdasarkan atribut kadar gula selama 116 hari giling pada tahun 2020. Hasil dari penelitian ini, atribut kualitas kadar gula terdiri dari rendemen sementara, pct brix, dan pct pol. Mutu dari kriteria visual tidak memiliki korelasi dengan atribut gula dan membuktikan tidak ditemukan bias antara mutu dari uji visual BBT dengan atribut kadar gula pada PT. XYZ. Jumlah klaster yang digunakan adalah empat klaster. Mutu A merupakan mutu terbaik karena memiliki nilai rata-rata dan range data tertinggi untuk setiap atribut kadar gula, kemudian mutu B, C, dan D. Urutan metode prediksi terbaik yang diuji yaitu, SVM Polinomial, SVM RBF, dan KNN. Performansi SVM lebih baik dibandingkan KNN jika terdapat set data yang kompleks dan berukuran besar dari atribut atau fitur prediktornya. Jenis kernel pada SVM mempengaruhi hasil akurasi, kernel membuat atribut atau fitur data asli dapat diproyeksikan ke dimensi yang lebih tinggi sehingga data tersebut dapat diklasifikasi dengan baik.**

Kata Kunci—Data Mining, Hierarchical K-Means Clustering, K-Nearest Neighbors, Manajemen Kualitas, Support Vector Machine.

I. PENDAHULUAN

BERDASARKAN kontribusi terhadap Produk Domestik Bruto (PDB), sektor pertanian, kehutanan, dan perikanan memiliki peran yang cukup besar yaitu sekitar 12,72% pada tahun 2019 serta merupakan urutan ketiga setelah sektor Industri Pengolahan dan sektor Perdagangan Besar dan Eceran; Reparasi Mobil dan Sepeda Motor. Salah satu sub sektor yang memiliki potensi besar yaitu perkebunan yang memiliki kontribusi sebesar 3,27% dan merupakan urutan pertama di sektor pertanian pada tahun 2019. Tebu merupakan salah satu komoditi perkebunan yang mempunyai peran strategis dalam perekonomian di Indonesia. Pemerintah

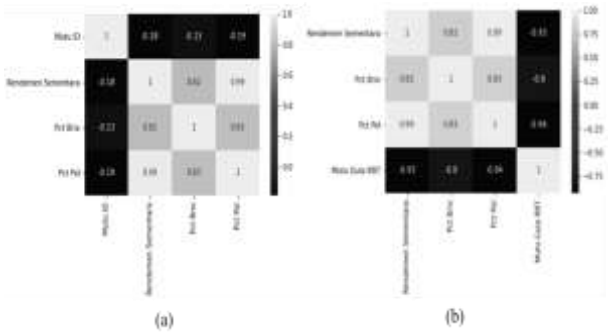
Gambar 1. Kriteria Uji Visual Kualitas Bahan Baku Tebu (BBT) pada perusahaan PT. XYZ.



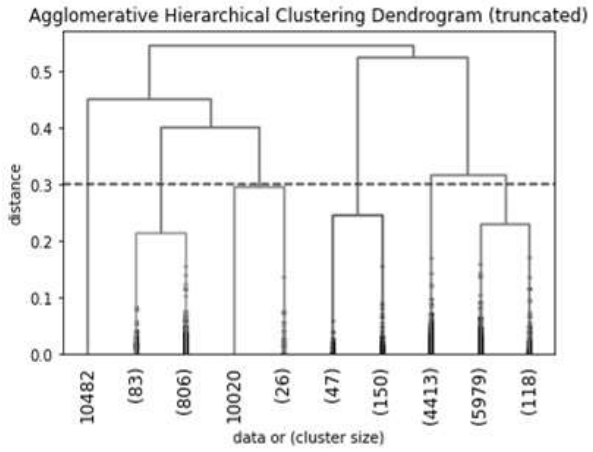
Gambar 2. Proses Budidaya Tebu dan Tebang Muat Angkut (TMA) Bagian Tanaman PT. XYZ.

berupaya agar Indonesia dapat mencapai swasembada gula sebagai salah satu langkah menuju Ketahanan Pangan Nasional.

Menurut Statistik Tebu Indonesia, konsep dan definisi Perusahaan Perkebunan adalah pelaku usaha perkebunan warga negara Indonesia atau badan hukum yang didirikan menurut hukum Indonesia dan berkedudukan di Indonesia yang mengelola usaha perkebunan dengan skala tertentu. Jawa Timur merupakan provinsi yang paling banyak memiliki produsen gula pada tahun 2019 yaitu 47,19%. Perkebunan tebu di Indonesia salah satunya diolah oleh PT. XYZ sebagai perusahaan yang berstatus sebagai Badan Usaha Milik Negara (BUMN). Salah satu Pabrik Gula (PG) milik PT. XYZ yang berlokasi di Jawa Timur memiliki mitra lebih dari 2000 petani tebu yang meliputi kabupaten Mojokerto, Jombang, dan Lamongan. Karena Bahan Baku



Gambar 3. (a) Hasil Uji Korelasi Mutu ID BBT; (b) Hasil Uji Korelasi Mutu Gula BBT Setelah Pengklasteran.

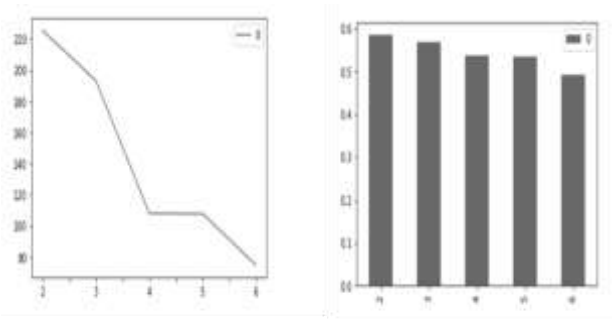


Gambar 4. Dendrogram *Average Linkage (Truncated*).

Tebu (BBT) yang diterima dan dibeli dari mitra petani tebu tergolong banyak, hal ini membuat varietas dan kualitas BBT yang diterima juga menjadi beraneka ragam. BBT yang baru diterima dari petani tebu akan dilakukan analisis visual oleh bagian *Quality Assurance (QA) on-farm* untuk dicek mutu BBT mulai dari A hingga E sebelum memasuki proses giling. Kategori mutu BBT A hingga E pada PT. XYZ memiliki kriteria masing-masing yang ditentukan dari analisis uji kriteria visual yang dapat dilihat pada Gambar 1.

Selain dilakukan analisis uji visual, PT. XYZ juga menguji kandungan kadar gula dari BBT yang diukur dengan menganalisis nira tebu atau sabut tebu. Analisis kualitas nira dan sabut tebu meliputi %brix, %pol, pH, Harkat Kemurnian, dan Nilai Nira Perahan Pertama (NNPP) [1]. Kualitas tebu juga dapat menentukan jumlah rendemen dan hablur yang didapatkan dari hasil *core sampler* yang dilakukan oleh bagian QA PT. XYZ. Hasil rendemen dan hablur tersebut digunakan untuk menentukan harga yang harus dibayarkan kepada petani tebu. Berikutnya, faktor penundaan giling dan varietas tanaman tebu juga mempengaruhi penyusutan bobot tebu, penurunan %pol atau kadar gula yang terkandung pada BBT yang akan digiling [1].

Analisis mutu BBT berdasarkan uji kriteria visual masih membuat adanya kemungkinan terjadi bias non-teknis atau penyimpangan yang dilakukan baik disengaja maupun tidak disengaja terhadap analisis kadar gula [2]. Petugas analisis bisa dimungkinkan memberikan penilaian berdasarkan kira-kira atau tebakan tanpa dianalisa terlebih dahulu misalnya, untuk mutu A memiliki nilai %brix 15-16, mutu B 14-15, mutu C, D, dan E memiliki nilai %brix 13-14 dan %pol 30-36 serta selalu menghasilkan angka rendemen 4-5 [2]. Untuk itu menganalisis BBT berdasarkan hasil uji kadar gula



Gambar 5. (a) Diagram Nilai SSE Untuk *Elbow Method*; (b) Diagram Nilai SWC Pada *Silhouette Method.*

diperlukan oleh PT. XYZ untuk mengetahui bagaimana penentuan kategori dan prediksi untuk klasifikasi dari kualitas BBT yang dikirim oleh mitra petani tebu sebelum masuk ke proses giling. Selain itu, penentuan kategori BBT berdasarkan uji kadar gula ini hanya berpengaruh kepada petani tebu dan bagian tanaman PT. XYZ sebagai evaluasi kinerja mereka dalam budidaya tanaman tebu dan proses Tebang Muat Angkut (TMA) sehingga diharapkan dapat meningkatkan keuntungan dalam bentuk bagi hasil antara petani tebu dan perusahaan PT. XYZ.

Penentuan kualitas BBT berdasarkan hasil analisis kadar gula tersebut bisa diatasi dengan pendekaan *data mining.* Pada beberapa penelitian, teknik prediksi *data mining* untuk klasifikasi menghasilkan akurasi yang tinggi untuk menentukan kualitas proses *injection molding* dengan menggunakan metode *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *General Classification and Regression Tree* (GC&RT) [3]. Pada penelitian-penelitian yang menggunakan metode penggabungan teknik pengklasteran seperti *K-Means Clustering* atau *Agglomerative Hierarchical Clustering* dengan metode prediksi untuk klasifikasi juga mendapatkan hasil yang baik karena teknik pengklasteran dapat meningkatkan akurasi model klasifikasi yang ditugaskan [4-5]. Dari hasil alternatif teknik pengklasteran dan prediksi untuk klasifikasi menggunakan pendekatan *data mining* tersebut, diharapkan pendekatan ini dapat membantu PT. XYZ dalam menentukan kelompok kualitas BBT berdasarkan kriteria analisis uji kadar gula dan model prediksi untuk klasifikasinya agar didapatkan perbaikan dari sistem penentuan mutu BBT dari uji kriteria visual.

Secara garis besar, dalam penelitian ini akan dilakukan perbaikan sistem penentuan kualitas atau mutu BBT yang terlepas dari uji kriteria visual menggunakan pendekatan *data mining* yaitu teknik pengklasteran dengan penggunaan metode *Hierarchical K-Means Clustering* untuk penentuan kelompok atau klaster mutu BBT dan karakteristik klasternya berdasarkan atribut dari uji kadar gula selama 1 tahun giling pada tahun 2020 yaitu 30 Mei hingga 22 September 2020 (selama 116 hari giling). Setelah dilakukan pengklasteran, dilanjutkan teknik prediksi untuk klasifikasi menggunakan metode *K-Nearest Neighbors* (KNN) dan *Support Vector Machine* (SVM). Pada tahap terakhir, metode klasifikasi tersebut dibandingkan dan dilakukan validasi sehingga akan didapatkan metode dengan hasil prediksi yang paling akurat agar bisa digunakan oleh PT. XYZ untuk memprediksi mutu BBT pada penerimaan di periode berikutnya. Selanjutnya

Tabel 1.
Ringkasan Hasil Pengklasteran k = 4 (4 Klaster)

Jumlah Data				Indeks Klaster	Mutu Gula BBT
11.624	1.858		15,98%	3	A
	3.720		32,00%	0	B
	5.842		50,26%	1	C
	204		1,75%	2	D
Aspek	Rendemen Sementara	Pct Brix	Pct Pol	Indeks Klaster	Mutu Gula BBT
Rata-rata	8,6835	15,7394	11,5606	3	A
	7,8109	14,6847	10,4113	0	B
	7,1176	13,4035	9,4967	1	C
	6,0407	12,9572	8,0381	2	D
Min Value	8,1500	13,5100	10,8500	3	A
	7,2100	11,8500	9,6100	0	B
	6,5200	10,0000	8,7100	1	C
	5,9200	10,0000	8,0000	2	D
Max Value	10,1900	20,7700	13,5000	3	A
	8,4200	17,7400	11,2300	0	B
	7,6200	17,2600	10,0800	1	C
	6,5400	15,9200	8,6600	2	D

dalam penelitian ini, metode pengumpulan dan pengolahan data dijelaskan pada Metode Penelitian, hasil pengolahan data dijelaskan pada Hasil dan Diskusi, dan Kesimpulan.

II. METODE PENELITIAN

A. Evaluasi Permasalahan dan Pra-Pengolahan Data

Evaluasi permasalahan di PT. XYZ yang didapatkan dari hasil wawancara, data riwayat dari bagian QA, dan penelitian-penelitian sebelumnya yang relevan dengan permasalahan di PT. XYZ. Data riwayat bagian QA PT. XYZ terdiri dari 65 kolom sejumlah 11.624 data yang didapatkan dari 30 Mei sampai 22 September 2020. Permasalahan yang diangkat yaitu penentuan kualitas BBT dari hasil analisis kadar gula *core sampler* yang digunakan oleh PT. XYZ, sehingga kolom yang digunakan sebagai atribut kualitas BBT yaitu: *1) Rendemen Sementara, 2) Pct Brix*, dan *3) Pct Pol*. Kolom "*Mutu ID*" akan digunakan untuk menguji kebiasan antara mutu dari uji visual dengan atribut analisis kadar gula pada BBT. Data pada "*Mutu ID*" berbentuk data kategori A hingga E, sehingga akan diubah menjadi data numerik 0 hingga 4. Set data dilanjutkan pada tahap pembersihan data. Hasil dari tahap pembersihan data ini tidak didapatkan *missing value* pada 11.624 data untuk setiap kolom atribut yang akan digunakan.

Setelah dilakukan pembersihan, set data riwayat perlu dilakukan transformasi data dengan *scaling* [6]. Pada proses *scaling* data akan dilakukan pada *range* 0 hingga 1 yang biasa disebut dengan Min-Max *scaling*.

$$\hat{X} = \frac{x - x_{min}}{x_{max} - x_{min}} \times (BA - BB) + BB \quad (1)$$

Pada persamaan (1), dimana BA atau batas atas adalah 1, sedangkan BB atau batas bawah adalah 0.

B. Uji Korelasi Atribut dengan Mutu Visual BBT

Uji korelasi ini dilakukan untuk mengecek kebiasan dari mutu visual dengan atribut dari hasil analisis kadar gula. Karena pada dasarnya kualitas dari kadar gula yang terkandung pada BBT tidak dapat dihubungkan dengan mutu dari uji visual BBT karena adanya faktor lain yang mempengaruhi seperti jenis varietas, umur BBT, waktu simpan, dan lain-lain [2], [7]. Secara umum, formulasi perhitungan koefisien korelasi (*r*) yaitu:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{(x - \bar{x})^2}\sqrt{(y - \bar{y})^2}} \quad (2)$$

Pada persamaan (2), dimana x dan y adalah data dari dua atribut yang ingin diuji korelasinya.

C. Tahap Pengklasteran Hierarchical K-Means Clustering

Pada tahap ini dikembangkan model *hierarchical k-means clustering* yaitu metode *hybrid* dari *agglomerative hierarchical clustering* dan *k-means clustering*. Metode ini digunakan karena kombinasi algoritma *hierarchical clustering* dan *k-means* menghasilkan hasil pengelompokan data yang lebih baik jika dibandingkan dengan *k-means* dalam semua pengujian [8].

1) Tahap 1: Metode Agglomerative Hierarchical Clustering

Pada *Agglomerative clustering* akan digunakan jenis *linkage clustering*: *complete, single, average, centroid,* dan *ward* [6]. Tahapan *Aggolmerative Hierarchical Clustering* sebagai berikut:

1. Untuk semua *N* data hitung matriks jarak antara tiap titik data menggunakan persamaa *Euclidean*
2. Kelompokkan setiap *N* titik data menjadi *N* klaster (klasternya sendiri) di *dendrogram*
3. Hitung jarak *linkage* antar klaster
4. Gabungkan dua klaster terdekat menjadi satu klaster
5. Ulangi hingga sampai tersisa satu klaster
6. Simpan *dendrogram*
7. Lakukan evaluasi *cophenetic correlation* untuk menentukan *linkage* terbaik dengan menggunakan persamaan *cophenet (c).*

$$c = \frac{\sum_{i<j}(Y_{ij} - y)(Z_{ij} - z)}{\sqrt{\sum_{i<j}(Y_{ij} - y)^2 \sum_{i<j}(Z_{ij} - z)^2}} \quad (3)$$

Pada persamaan (3), dimana *c* adalah koefisien korelasi *cophenet*, Z_{ij} adalah jarak *linkage* antar klaster, Y_{ij} adalah jarak asli antar objek (*dissimilarity*) pada *dendrogram*, dan n_j adalah banyaknya objek atau anggota dalam klaster ke-j

2) Tahap 2: Metode K-Means Clustering

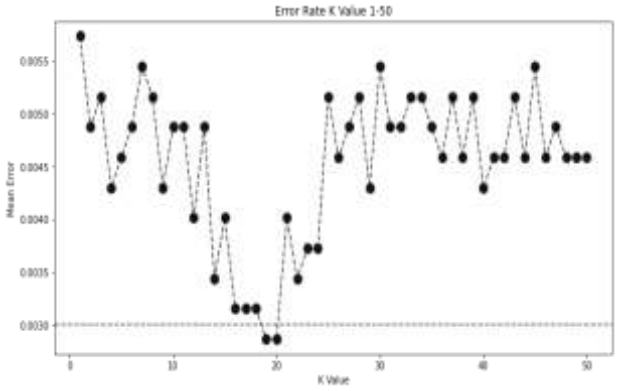
Pada penggunaan metode *k-means clustering* akan ditugaskan untuk jumlah dan jenis klaster yang didapatkan dari hasil evaluasi *Agglomerative Hierarchical Clustering* [9]. Tahapan *K-Means Clustering* sebagai berikut:

1. Tentukan jumlah *k* klaster yang ditugaskan dari hasil AHC
2. Alokasikan data ke dalam klaster sesuai hasil AHC
3. Hitung *centroid* inisial dari data yang ada di masing-masing klaster AHC. Lokasi *centroid* inisial setiap klaster diambil dari rata-rata (*mean*) semua nilai data pada setiap atribut [6].

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \quad (4)$$

Pada persamaan (4), dimana *M* menyatakan jumlah data dalam sebuah klaster, *i* menyatakan fitur/atribut ke-*i* dalam sebuah klaster.

4. Hitung jarak setiap data ke setiap *centroid* menggunakan persamaan *Euclidean* [6].



Gambar 6. Hasil Rata-rata *Error* Nilai k= 1 hingga k=50 dari Proses KNN.

$$E = \sum_{j=i}^{k} \sum_{i=1}^{n_j} \left|\left| x_i^j - c_j \right|\right|^2 \quad (5)$$

Pada persamaan (5), dimana x_i^j adalah objek ke-i pada klaster ke-j, c_j adalah pusat klaster (*centroid*) ke-j, k adalah jumlah klaster, dan n_j adalah banyaknya objek atau anggota dalam klaster ke-j

5. Alokasikan kembali data ke dalam klaster yang memiliki jarak dengan *centroid* terdekat
6. Hitung *centroid* yang baru dari data yang sudah dialokasikan kembali
7. Ulangi tahap 4 sampai 6 hingga lokasi *centroid* tidak berubah lagi (konvergen)
8. Lakukan evaluasi pertama, yaitu *elbow method.* Pada evaluasi model dengan *Elbow Method*, diperoleh jumlah *k* klaster yang optimal ketika selisih nilai SSE dari *k* sebelumnya berkurang secara signifikan dan membentuk *elbow*.

$$SSE = \sum_{i=1}^{k} \sum_{x \in D_i} \left|\left| x - m_i \right|\right|^2 \quad (6)$$

Pada persamaan (6), dimana dalam suatu klaster D_i, vektor rata-rata m_i adalah nilai *centroid* yang paling baik untuk mewakili data-data dalam klaster D_i tersebut, yaitu yang meminimalkan nilai dari SSE [6].

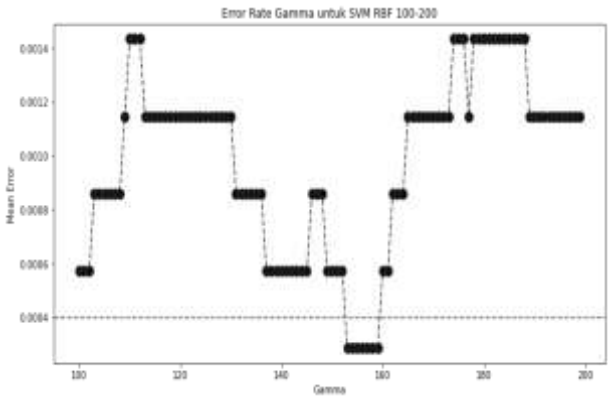
9. Lakukan evaluasi kedua, yaitu *silhouette method.* Metode ini menghitung koefisien *Silhouette* $s(i)$ dari setiap titik data untuk mengukur seberapa mirip suatu titik data tersebut dengan klasternya sendiri dibandingkan dengan klaster yang lain.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad (7)$$

Pada persamaan (7), dimana nilai *silhouette* $s(i)$ dicari dengan menghitung rata-rata jarak objek ke-i ke sesama objek dalam satu klaster $a(i)$ dan rata-rata objek ke-i ke objek di klaster yang lain $b(i)$.

$$SWC = \frac{1}{N} \sum_{i=1}^{N} s(i) \quad (8)$$

Pada persamaan (8), dimana parameter *Silhouette Width Criterion (SWC)* adalah hasil dari rata-rata nilai *silhouette* $s(i)$ dari setiap objek ke-i. Jumlah klaster yang optimal adalah yang memberikan nilai SWC paling tinggi.



Gambar 7. Hasil Rata-rata *Error* Parameter *Gamma* 100 hingga 200 dari Proses SVM RBF.

D. Pembagian Data Klasifikasi

Pada tahap ini data hasil pengklasteran akan dibagi sebagai data *training* dan data *testing* sebelum masuk ke proses klasifikasi. Rasio pembagian yang digunakan yaitu 70% untuk data *training* dan 30% untuk data *testing.* Pembagian data klasifikasi dilakukan sebanyak 5 kali (N=5) untuk tahap validasi pada akhir proses prediksi.

E. Tahap Pengembangan Model Prediksi Klasifikasi

Hasil dari pengklasteran untuk penentuan label atau kelas dan karakteristik BBT yaitu bersifat data kategori dan bukan data numerik. Sehingga untuk metode prediksi KNN dan SVM ini akan digunakan sebagai model klasifikasi. Untuk model klasifikasi yang digunakan yaitu *multiclass* karena banyaknya jumlah label yang digunakan pada tahap ini bergantung dari hasil evaluasi pada tahap *clustering*.

1) K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) adalah salah satu metode klasifikasi *supervised* yang umum digunakan dan termasuk dalam kategori klasifikasi *non-parametric.*

Misalkan *k* menyatakan jumlah data atau tetangga terdekat dari data *training* terhadap data *testing*, maka tahapan *k-nearest neighbors* sebagai berikut:

a. *Input* set data *training* beserta labelnya, data *testing*, dan nilai *k* yang diuji
b. Untuk semua data *testing i* hitung jaraknya ke setiap data *training* menggunakan jarak *Euclidean*
c. Tentukan *k*-data *training* yang jaraknya paling dekat dengan data *testing i*
d. Periksa label dari *k* data tersebut
e. Tentukan label dengan frekuensi paling banyak
f. Masukan data *testing i* tersebut ke kelas dengan frekuensi yang paling banyak
g. Ulangi langka 3 sampai 6 untuk semua data *testing* yang lainnya

Pada penentuan nilai *k* tetangga terdekat terbaik yang akan digunakan, didapat dari pengujian rata-rata *error* dari setiap *k* yang diuji yaitu k=1 hingga k=50. Nilai *k* yang digunakan yaitu yang memiliki nilai rata-rata *error* terkecil.

2) Support Vector Machine (SVM)

Pada metode SVM, fungsi kernel yang digunakan adalah untuk data non-linear yaitu, Polinomial dan *Radial Basis Function* (RBF). Polinomial merupakan salah satu jenis kernel yang digunakan karena sifatnya yang sederhana. RBF

adalah salah satu kernel paling kuat dan umum digunakan di SVM. Pada metode SVM dengan kernel RBF dan polinomial akan digunakan parameter pinalti *C* =1.

Variabel dan parameter:

$x = \{x_1, x_2, \dots, x_l\}$: $sample\ training$

$y = \{y_1, y, \dots, y_l\}$C $\{\pm 1\}$: $label\ data\ training$

$kernel$:fungsi kernel

par :parameter

C :konstanta pinalti *cost slack*

$\alpha = [\alpha_1, \alpha_2, \dots, \alpha_l]$:*Langrange multiplier*, dan bias *b*

a. Hitung matriks Kernel K berdasarkan fungsi dan parameter kernel yang ditugaskan.
b. Tentukan pembatas untuk programa kuadratik (*QP*)
c. Tentukan fungsi tujuan programa kuadratik (*QP*) menggunakan formula SVM *dual problem.*

$$\max \alpha \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j\, y_i y_j {x_i}^T x_j$$

$$subject\ to: \quad (9)$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0; C_i \geq \alpha_i \geq 0\ untuk\ i = 1,2, \dots, l$$

Pada persamaan (9), dimana ${x_i}^T x_j$ dapat diganti menggunakan fungsi kernel *k* yang ditugaskan.

d. Selesaikan masalah QP dan temukan solusi α_i dan *b* menggunakan persamaan (10) dan (11).

$$b = y_i - w^T x_i \quad (10)$$

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \quad (11)$$

e. Gunakan parameter yang didapatkan dari langkah 4 untuk melakukan prediksi menggunakan formula prediksi.

$$y_{prediksi} = sign(w^T x + b)$$

$$y_{prediksi} = sign(\sum_{i=1}^{l} \alpha_i y_i (x_i x) + b) \quad (12)$$

Pada persamaan (12), dimana $(x_i x)$ adalah perkalian antara data *testing* dengan *support vectors*.

Pada kernel RBF, paramater yang akan dicari yaitu nilai *gamma* dari 1-500 dengan menggunakan hasil rata-rata *error* terkecil. Pada kernel polinomial terdapat dua parameter yaitu, *gamma* dan *degree*. Untuk dua parameter ini akan dilakukan dua percobaan berbeda, antara lain:

1) Percobaan 1: Parameter nilai *gamma* yang akan digunakan yaitu *gamma* dari hasil kernel RBF sebelumnya, sedangkan parameter kedua yaitu derajat atau *degree* polinomial akan dicari menggunakan rata-rata *error* terkecil antara d=1 hingga d=15.
2) Percobaan 2: Parameter derajat atau *degree* yang akan digunakan yaitu 3 berdasarkan *default* dari Python, sedangkan parameter *gamma* akan dicari menggunakan rata-rata *error* terkecil.

Dari kedua percobaan berikut akan dipilih salah satu yang memiliki nilai *error* terkecil. Berikut merupakan algoritma SVM untuk klasifikasi [6].

F. Tahap Evaluasi Model Prediksi Klasifikasi

Metode validasi yang umum digunakan dalam teknik prediksi untuk klasifikasi kategori adalah *confusion matrix.* *Confusion matrix* adalah penggunaan tabel untuk memvisualisasikan kinerja model prediksi untuk kategori atau kelas [6]. Pada tahap ini akan dilakukan uji validasi untuk masing-masing teknik prediksi dari hasil data *training* dan data *testing.* Hasil klasifikasi dalam bentuk kategori dan bukan numerik, sehingga teknik validasi yang akan diimplementasikan yaitu *confusion matrix* untuk klasifikasi *multi-class*. Teknik validasi ini akan digunakan ukuran akurasi, presisi, dan sensitifitas (*recall*) untuk setiap hasil prediksi dari masing-masing metode klasifikasi KNN dan SVM dengan percobaan pembagian atau *split* data *training* dan *testing* sebanyak 5 kali.

Pada bagian ini akan ditinjau sensitivitas model prediksi dari pengaruh besarnya jumlah data dari pembagian atau *split* data *training* 70% dan *testing* 30% terhadap akurasi model prediksi KNN, SVM RBF, dan SVM Polinomial. Perhitungan uji sensitivitas besar data untuk fraksi data yang diuji yaitu 10% hingga 100% dari 11.624 jumlah data.

G. Tahap Implementasi Model

Model prediksi terbaik yang memiliki nilai akurasi, presisi, dan *recall* yang tertinggi serta memiliki hasil uji sensitivitas terbaik dari jumlah data yang diuji. Model prediksi terbaik tersebut dapat digunakan sebagai model perancangan sistem database baru perusahaan dengan mengintegrasikan ke SAP perusahaan. Semua pengolahan data dan uji statistik dalam penelitian ini dieksekusi menggunakan bahasa pemrograman *Python 3.0* dan *software* JUPYTER yang bersifat *open-source*.

III. HASIL DAN DISKUSI

A. Hasil Sistem Penentuan Mutu BBT dari Uji Kadar Gula

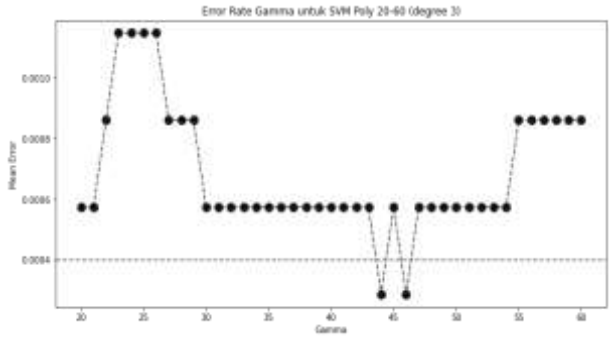
Sistem penentuan mutu BBT dari uji kadar gula digunakan sebagai evaluasi kualitas tanaman tebu yang diterima dari petani tebu yang dimiliki. Menurut P3GI, dalam menyelesaikan permasalahan kualitas pasok tanaman tebu adalah dengan melakukan analisa tebu secara langsung menggunakan uji *core sampler* [10]. Penerimaan tebu adalah tanggung jawab bagian tanaman perusahaan. Proses bagian tanaman terdapat pada Gambar 2.

Bagian tanaman memiliki tugas utama untuk menyediakan BBT dan bertanggung jawab secara keseluruhan terhadap pemenuhan BBT baik jumlah dan mutunya. Bagian tanaman secara umum dibagi menjadi dua, yaitu budidaya tanaman tebu dan Tebang Muat Angkut (TMA). Budidaya tanaman tebu yaitu pengelolaan tebu mulai dari pembibitan, persiapan lahan, penanaman, dan pemeliharaan. TMA terdiri dari analisis umur tebu, analisis kemasakan tebu, jadwal tebang, dan penentuan kriteria Masak, Bersih, Segar (MBS).

Petani dapat meningkatkan kualitas tanaman tebu yang mereka tanam dengan bantuan bagian tanaman PT. XYZ yaitu dalam bentuk konsultasi atau perbaikan program pembinaan budidaya tanaman tebu yang sudah dimiliki. Peningkatkan kualitas tanaman tebu atau BBT berdasarkan uji kadar gula ini dapat meningkatkan keuntungan dalam bentuk bagi hasil antara petani tebu dan perusahaan.

B. Hasil Uji Korelasi Mutu BBT

Uji korelasi dilakukan untuk mengecek kebiasan dari mutu visual dengan atribut kualitas kadar gula. Atribut kualitas yang didapatkan dari kolom data riwayat PT. XYZ dan



Gambar 8. Hasil Rata-rata *Error* Parameter *Gamma* 20 hingga 60 dari Proses SVM Polinomial (Percobaan 2).

prosedur QA yaitu rendemen (rendemen sementara), %brix (pct brix), dan %pol (pct pol). Rendemen adalah fungsi dari kualitas tebu dan efisiensi pabrik, %brix adalah zat padat kering terlarut dalam larutan (gr/100gr larutan) yang dihitung sebagai sukrosa, dan %pol adalah jumlah gula (gr) yang terlarut dalam 100 gram larutan yang mempunyai kesamaan putaran optik dengan sukrosa murni. Atribut kualitas dari kadar gula yang terkandung dalam BBT tidak dapat dihubungkan dengan mutu dari uji visual BBT karena adanya faktor lain yang mempengaruhi hasil kualitas dari kadar gula yang terkandung dalam BBT seperti jenis varietas, umur BBT, waktu simpan, dan lain-lain [2], [7].

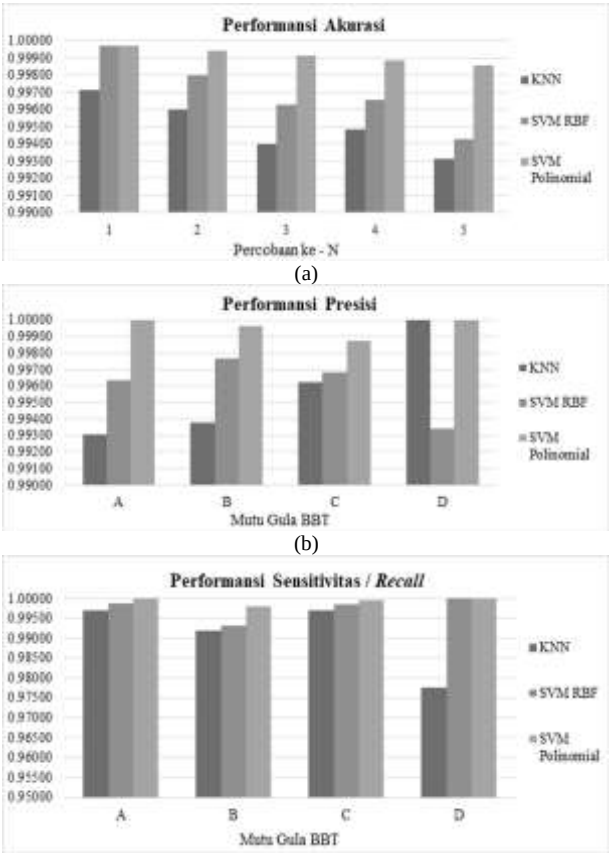
Pada hasil matriks Gambar 3(a) dari uji korelasi, *Mutu ID* yang berasal dari uji kriteria visual BBT memiliki korelasi yang rendah terhadap atribut kadar gula. Hal ini menjelaskan bahwa *Mutu ID* tidak memiliki korelasi dengan atribut gula yang diuji dan membuktikan bahwa tidak ditemukan bias antara mutu dari uji visual BBT dengan atribut kadar gula pada PT. XYZ. Hasil uji korelasi yang dilakukan dari hasil pengklasteran pada Gambar 3(b) menunjukkan bahwa *Mutu Gula BBT* memiliki hasil korelasi yang tinggi terhadap atribut gula. Hal ini menjelaskan bahwa mutu gula BBT yang didapatkan dari teknik pengklasteran memiliki korelasi dengan atribut kadar gula yang dimiliki.

C. Hasil Pengklasteran Hierarchical K-Means Clustering

Hasil tahap pertama yaitu *agglomerative hierarchical clustering,* penentuan *linkage clustering* terbaik berdasarkan nilai *cophenet* (*c*). Jenis *linkage clustering* yang diuji yaitu *single, complete, average, centroid,* dan *ward*. Dari perhitungan *cophenet* (*c*), didapatkan bahwa metode *linkage* terbaik yaitu *average* karena memiliki nilai *c* tertinggi 0,79258. Kemudian akan dibentuk pohon klaster atau dendrogram untuk menentukan pengelompokan datanya, jumlah klaster yang dibentuk yaitu k=2 hingga k=6 berdasarkan Gambar 4.

Hasil rata-rata pengelompokan data dari alternatif jumlah klaster k tersebut akan digunakan sebagai inisial *centroid* yang berbentuk *array* untuk tahap *k-means clustering.* Tahap kedua yaitu *k-means clustering,* dilakukan evaluasi menggunakan nilai SSE untuk metode *elbow method* dan nilai SWC untuk *Silhouette* untuk k=2 hingga k=6.

Pada Gambar 5(a) didapatkan bahwa perubahan nilai SSE yang signifikan dan membentuk *elbow* pada k=4. Pada Gambar 5(b), nilai SWC untuk alternatif jumlah klaster k=2 hingga k=6 mengalami penurunan untuk setiap alternatifnya sehingga SWC terbaik yaitu k=2 karena memiliki nilai SWC tertinggi yaitu 0,5849. Dari hasil kedua metode evaluasi



Gambar 9. (a) Performansi Akurasi Model Untuk 5 Kali Percobaan; (b) Performansi Presisi Model Terhadap Mutu Gula BBT; (c) Performansi Sensitivitas/*Recall* Model Terhadap Mutu Gula BBT.

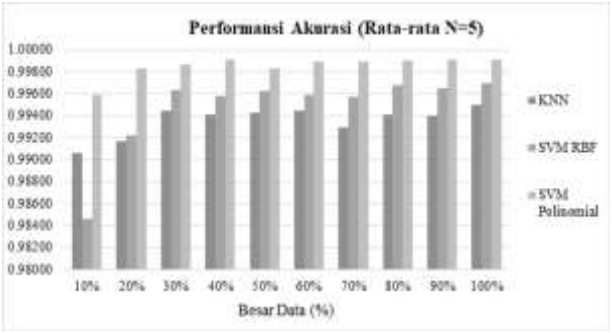
tersebut akan diambil jumlah klaster k=4 berdasarkan hasil evaluasi *elbow method* dengan nilai SWC yang dimiliki yaitu 0,5368 dan memiliki selisih 0,048 dari hasil SWC k=2. Nilai selisih 0,048 antara k=2 dengan k=4 masih dapat ditoleransi.

Penentuan karakteristik hasil pengklasteran dengan k=4 disajikan pada Tabel 1. Jumlah data tiap klaster secara berurutan dari yang terbanyak yaitu mutu C, B, A, dan D dengan persentase-nya yaitu 50,26%, 32,00%, 15,98%, dan 1,75% dari 11.624 data riwayat yang dimiliki. Pada hasil rata-rata, nilai minimal, dan nilai maksimal data di setiap klaster, semua kategori mutu pada klaster memiliki nilai berurutan dari yang tertingi ke terendah untuk semua atribut kadar gula yang dimiliki dari mutu A hingga mutu D.

Pada Tabel 1 menunjukkan bahwa mutu A merupakan mutu terbaik karena memiliki nilai rata-rata dan *range* data tertinggi untuk setiap atribut kadar gula yang dimiliki dibandingkan mutu yang lain. Setelah mutu A, terdapat mutu B, C, dan yang terburuk kualitas gulanya yaitu mutu D. Pada hasil karakteristik pengklasteran ini dapat dikatakan baik karena untuk setiap karakteristik klasternya mulai dari mutu A hingga mutu D memiliki nilai dari aspek rata-rata dan *range* data yang berurutan dari yang tertinggi ke terendah untuk semua atribut kadar gula yang diuji [8], [11].

D. Hasil Prediksi Model Klasifikasi KNN dan SVM

Pada bagian ini data hasil pengklasteran akan diubah menggunakan indeks baru untuk "Mutu Gula BBT" yaitu {A, B, C, D} menjadi {0, 1, 2, 3}. Data klaster akan dibagi menjadi variabel x sebagai fitur atau atribut dan variabel y sebagai target atau luaran untuk teknik prediksi. Kemudian



Gambar 10. Hasil Perhitungan Sensitivitas Besar Data 10% hingga 100% dari 11.624 Jumlah Data Berdasarkan Performansi Akurasi (N=5).

akan dilakukan pembagian data dari hasil pengklasteran dengan perbandingan 70% data *training* dan 30% data *testing* yang dibagi tiap klasternya. Pembagian data klasifikasi dilakukan sebanyak 5 kali (N=5) untuk tahap validasi dan uji sensitivitas besar data pada akhir proses prediksi.

1) K-Nearest Neighbors (KNN)

Pada tahap pengembangan model prediksi KNN dan penentuan nilai k digunakan data *training* dan data *testing* pada pembagian data pertama. Nilai k=1 hingga k=50 diuji untuk mendapatkan nilai k tetangga terdekat dengan rata-rata *error* terkecil pada Gambar 6.

Pada Gambar 6, rata-rata *error* terkecil dari nilai k tetangga terdekat yaitu k=19 dan k=20 memiliki hasil yang sama. Model prediksi KNN dengan nilai k dari salah satu hasilnya, yaitu k=20 digunakan untuk melatih data *training* dan dilanjutkan untuk memprediksi data *testing.* Hasil variabel y dari prediksi KNN akan divalidasi dengan hasil variabel y dari data *testing* menggunakan *confusion matrix.*

2) Support Vector Machine (SVM) dengan Kernel RBF

Pada model prediksi SVM dengan jenis kernel *Radial Basis Function* (RBF) akan dilakukan penyesuaian parameter *gamma.* Penentuan nilai *gamma* yang akan digunakan akan dipilih berdasarkan nilai *error* terkecil. Nilai *gamma* yang diuji yaitu 1 hingga 500 dan didapatkan *gamma* terkecil terletak pada *range* 100 hingga 200. *Range* 100 hingga 200 diuji dan didapatkan *gamma* dengan *error* terkecil pada Gambar 7.

Pada Gambar 7 didapatkan bahwa *error* terkecil terletak antara *gamma* 153 hingga 159. Karena memiliki nilai *error* yang sama antara 153 hingga 159, maka dapat diambil salah satu nilai *gamma* untuk digunakan yaitu 155. *Gamma* 155 digunakan untuk melatih data *training* dan dilanjutkan untuk memprediksi data *testing.* Hasil variabel y dari prediksi SVM RBF akan divalidasi dengan hasil variabel y dari data *testing* menggunakan *confusion matrix.*

3) Support Vector Machine (SVM) dengan Kernel Polinomial

Pada kernel polinomial terdapat dua parameter yaitu, *gamma* dan *degree* atau derajat polinomial. Untuk dua parameter ini akan dilakukan dua percobaan berbeda. Pada percobaan 1, parameter nilai *gamma* yang akan digunakan yaitu *gamma* dari hasil kernel RBF sebelumnya (*gamma*=155), sedangkan parameter kedua yaitu derajat atau *degree* polinomial akan dicari menggunakan rata-rata *error* terkecil antara d=1 hingga d=50. Dari pengujian tersebut didapatkan *error* terkecil pada *range* 1 hingga 10. Pada percobaan 2, parameter derajat atau *degree* yang akan digunakan yaitu d=3, sedangkan parameter *gamma* akan dicari menggunakan rata-rata *error* terkecil. Nilai *gamma* yang diuji yaitu 1 hingga 500 dan didapatkan *gamma* terkecil terletak pada *range* 20 hingga 60. *Range* 20 hingga 60 kemudian diuji dan didapatkan *gamma* dengan *error* terkecil pada Gambar 8.

Pada kedua percobaan tersebut, didapatkan bahwa *error* yang dihasilkan oleh percobaan 2 lebih kecil. Karena memiliki nilai *error* yang sama antara *gamma* 44 dan 46, maka dapat diambil salah satu nilai *gamma* untuk digunakan sebagai pembangkit model prediksi SVM yaitu 44. Parameter *degree*=3 dan *gamma*=44 digunakan untuk melatih data *training* dan dilanjutkan untuk memprediksi data *testing.* Hasil variabel y dari prediksi SVM Polinomial akan divalidasi dengan hasil variabel y dari data *testing* menggunakan *confusion matrix.*

4) Validasi dan Sensitivitas Besar Data Pada Model Prediksi

Validasi model prediksi akan digunakan hasil dari *confusion matrix* dengan 5 kali (N=5) percobaan *split* data *training* dan *testing* secara *random*. Hasil performansi ukuran dari *confusion matrix* disajikan pada Gambar 9.

Hasil ketiga ukuran performansi tersebut dapat disimpulkan bahwa metode prediksi SVM Polinomial memiliki performansi yang paling bagus untuk diimplementasikan di PT. XYZ. Metode SVM RBF terletak pada urutan kedua dan terakhir yaitu KNN. Hasil ketiga ukuran performansi untuk metode KNN, SVM RBF, dan SVM Polinomial pada Gambar 9 cenderung tinggi yaitu antara 97% hingga 100%, hal ini dikarenakan adanya pengaruh dari hasil pengklasteran yang dilakukan sebagai penentuan kelas untuk kategori mutu gula BBT. Dari hasil pengklasteran, "Mutu Gula BBT" memiliki korelasi yang tinggi dengan atribut atau fitur prediktor yang dimiliki, sehingga metode prediksi yang ditugaskan juga menjadi lebih mudah untuk melatih data *training* dan memprediksi data *testing.*

Hasil uji sensitivitas untuk besar data yang diuji dengan 5 kali percobaan disajikan pada Gambar 10. Besar data 20% (2.325 data) hingga 100% (11.624 data), SVM Polinomial memiliki rata-rata akurasi tertinggi dan yang terendah yaitu KNN. Pada besar data 10% (1.162 data), SVM Polinomial juga menghasilkan akurasi tertinggi, tetapi akurasi terendah terdapat pada SVM RBF. Pada besar data 10% (1.162 data) hingga 30% (3.487 data), ditemukan pada beberapa percobaan hasil akurasi KNN masih dapat lebih tinggi atau sama terhadap SVM RBF, tetapi tidak ditemukan hasil akurasi KNN yang lebih tinggi dari SVM Polinomial. Hal ini menunjukkan bahwa selain dari jumlah data yang digunakan, jenis kernel juga dapat mempengaruhi hasil akurasi pada model SVM. Fungsi kernel pada metode SVM digunakan supaya atribut atau fitur data asli dapat diproyeksikan ke dimensi yang lebih tinggi sehingga data tersebut dapat diklasifikasi dengan baik menggunakan SVM [6].

Dari hasil ini dapat dikatakan bahwa SVM Polinomial dan RBF lebih unggul dibandingkan KNN, hal ini disebabkan adanya pengaruh ukuran set data untuk data *training* dan *testing* yang besar serta jenis kernel yang digunakan untuk pembentukan *classifier* pada SVM sehingga mampu menangani model prediksi yang kompleks dari atribut atau

fitur prediktornya [12]. Metode KNN tidak memiliki parameter dan hanya ditentukan berdasarkan nilai k tetangga terdekatnya dan persamaan jarak yang digunakan, sehingga jika terdapat set data yang kompleks dan berukuran besar dari atribut atau fitur prediktor, KNN tidak dapat memprediksi dengan baik jika dibandingkan dengan SVM yang menggunakan metode kernel tertentu.

IV. KESIMPULAN

Penelitian ini menjelaskan bahwa mutu dari kriteria visual tidak memiliki korelasi dengan atribut gula yang diuji sehingga membuktikan tidak ditemukan bias antara mutu dari uji visual BBT dengan atribut kadar gula. Perbaikan sistem penentuan kualitas atau mutu BBT menggunakan teknik pengklasteran *Hierarchical K-Means Clustering* menghasilkan bahwa jumlah klaster yang optimal adalah 4 klaster berdasarkan hasil evaluasi. Empat klaster tersebut yaitu, mutu A (15.98%), B (32%), C (50.26%), dan D (1.75%). Hasil karakteristik pengklasteran ini dapat dikatakan baik, mulai dari mutu A hingga D memiliki rata-rata dan *range* data yang berurutan dari yang tertinggi ke terendah untuk semua atribut kadar gula.

Pada klasifikasi dari data hasil pengklasteran berdasarkan hasil dari confusion matrix, metode prediksi terbaik yaitu: *SVM Polinomial*, *SVM RBF*, dan *KNN*. Pada uji sensitivitas, didapatkan hasil akurasi *KNN* masih dapat lebih tinggi atau sama terhadap *SVM RBF*, sedangkan *SVM Polinomial* menghasilkan akurasi tertinggi untuk semua besar data yang diuji. Set data yang kompleks dan berukuran besar dari atribut atau fitur prediktor, metode KNN tidak dapat memprediksi dengan baik jika dibandingkan dengan SVM yang menggunakan metode kernel tertentu. Penggunaan kernel pada SVM ini digunakan supaya atribut atau fitur data asli dapat diproyeksikan ke dimensi yang lebih tinggi sehingga data tersebut dapat diklasifikasi dengan baik.

Hasil dari penelitian ini diharapkan PT. XYZ menganalisis kualitas BBT berdasarkan hasil uji kadar gula sebelum masuk ke proses giling dengan melakukan integrasi sistem antara *database* bagian *Quality Assurance* (QA) pada *System Application and Processing* (SAP) perusahaan dengan *machine learning* (ML) untuk bahasa pemrograman *python* agar dapat dilakukan prediksi mutu gula BBT. Analisis kualitas BBT ini digunakan sebagai bahan evaluasi kinerja petani tebu agar dapat meningkatkan kualitas tanaman tebu yang ditanam dengan bantuan bagian tanaman dalam bentuk konsultasi atau perbaikan program pembinaan budidaya tanaman tebu yang dimiliki. Peningkatkan kualitas tanaman tebu atau BBT berdasarkan uji kadar gula ini dapat meningkatkan keuntungan dalam bentuk bagi hasil antara petani tebu dan perusahaan. Penelitian ini hanya berdasarkan tiga jenis atribut kadar gula, sehingga PT. XYZ disarankan untuk bisa menambahkan sistem uji *core sampler* menggunakan *NIR liquid* agar didapatkan hasil atribut seperti pH, gula reduksi, dan NPP.

DAFTAR PUSTAKA

[1] A. D. Kuspratomo, B. Burhan, and M. Fakhry, "Pengaruh varietas tebu, potongan dan penundaan giling terhadap kualitas nira tebu," *Agrointek*, vol. 6, no. 2, pp. 123–132, 2012.

[2] L. Manalu, "Studi kasus penentuan rendemen tebu di pabrik gula BUMN," *J. Keteknikan Pertan.*, vol. 20, no. 1, pp. 1–8, 2006.

[3] E. V Ramana, S. Sapthagiri, and P. Srinivas, "Data mining approach for quality prediction of injection molding process through statistica SVM, KNN and GC & RT techniques," *Inject. molding*, vol. 1, p. 9, 2016.

[4] Z. Yong, L. Youwen, and X. Shixiong, "An improved KNN text classification algorithm based on clustering," *J. Comput.*, vol. 4, no. 3, pp. 230–237, 2009.

[5] N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, "Improving classification accuracy using clustering technique," *Bull. Electr. Eng. Informatics*, vol. 7, no. 3, pp. 465–470, 2018.

[6] M. K. F. Hanka and B. Santosa, "Analisis Kualitas Bahan Baku Tebu Melalui Teknik Pengklasteran Dan Klasifikasi Kadar Gula Sebelum Giling (Studi Kasus Pabrik Gula PT. XYZ)," Institut Teknologi Sepuluh Nopember, 2021.

[7] H. YUHARTONO, "Faktor-Faktor yang Mempengaruhi Produksi Hablur Tebu Pabrik Gula Tersana Baru," Universitas Gadjah Mada, 2009.

[8] T. Alfina and B. Santosa, "Analisa Perbandingan Metode Hierarchical Clustering, K-Means Dan Gabugan Keduanya Dalam Membentuk Cluster Data (Studi Kasus : Problem Kerja Praktek Jurusan Teknik Industri Its)," Institut Teknologi Sepuluh Nopember Surabaya, 2012.

[9] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations," *Acm Sigmod Rec.*, vol. 31, no. 1, pp. 76–77, 2002.

[10] B. E. Santoso, "Rendemen: Definisi, Prosedur dan Kaitannya dengan Kinerja Pabrik," *Pusat Penelitian Perkebunan Gula Indonesia*. Pusat Penelitian Perkebunan Gula Indonesia, Pasuruan, 2002.

[11] T.-S. Chen *et al.*, "A Combined K-Means and Hierarchical Clustering Method for Improving The Clustering Efficiency of Microarray," in *2005 International symposium on intelligent signal processing and communication systems*, 2005, pp. 405–408.

[12] J. S. Raikwal and K. Saxena, "Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set," *Int. J. Comput. Appl.*, vol. 50, no. 14, 2012.