P-ISSN: 2962-3227 | E-ISSN: 2962-3235
Prosiding Seminar Hi-Tech (Vol. 1, No. 1, 2022)
https://ejournal.unuja.ac.id/index.php/hitech

Komparasi Metode KNN dan C4.5 dalam Klasifikasi Data Mutu Padi Organik

M. Erfan Rianto

Prodi Teknik Informatika, Universitas Nurul Jadid, Probolinggo, Indonesia
{erfankhece97@gmail.com}

Abstrak. Tanaman padi adalah tanaman semusim yang berumur pendek. Jenis varietas padi di indonesia sangatlah banyak dan bervariasi sehingga menjadi pokok pangan utama masyarakat indonesia (Adhy, 2021). Pada penelitian ini akan di bahas mengenai penerapan metode data mining. Bertujuan untuk membandingkan hasil dari kedua metode klasifikasi pada tanaman padi organik tersebut. Adapun metode yang dipakai yaitu Algoritma K-nearest Neighbors (KNN) dan C4.5 dengan menggunakan bahasa pemrograman python (Rachman et al., n.d.-a; Sari & Hasibuan, 2019). Data mutu padi organik kami peroleh dari dinas pertanian Bondowoso melalui link dari ***kaggle.com***. Kemudian hasil dari perbandingan kedua algoritma akan dijadikan acuan untuk klasifikasi. Kesimpulan dari prosiding ini adalah membandingkan hasil dari kedua algoritma tersebut.

Katakunci Kata kunci: *C4.5; Klasifikasi; K-nearest neighbors; Mutu padi organik*

Abstract. Rice plants are short-lived annuals. The types of rice varieties in Indonesia are very many and varied so that they become the main food staple of the Indonesian people. In this study, we will discuss the application of the data mining method. Aims to compare the results of the two classification methods on organic rice plants. The methods used are the K-nearest Neighbors (KNN) algorithm and C4.5 using the python programming language. We obtained organic rice quality data from the agriculture office Bondowoso via a link from ***kaggle.com.*** Then the results of the comparison of the two algorithms will be used as a reference for classification. The conclusion of this proceeding is to compare the results of the two agoritemes.

Keywords: *C4.5; Classification; K-nearest neighbors; Organic rice quality*

Pendahuluan

Padi adalah tanaman yang banyak di budidayakan oleh masyarakat indonesia. Tanaman padi atau dengan nama latinnya (Oryza sativa. L) adalah tanaman yang sangat subur di daerah tropis. Bukti sejarah di Cina Utara, membuktikan bahwa tananan tersebut di Asia sudah ada sejak 700 tahun lalu. Beberapa daerah rule diduga menjadi daerah asal tanaman padi adalah Bharat Utara dan Tamburlaine, Banglades Utara dan daerah bagian Birma,

Thailand, Laos Vietnam dan Cina Selatan. Meningkatkan hasil padi dengan cara menambah dosis pupuk anorganik dan pestisida bisa mengakibatkan kerusakan pada tanah. Kondisi seperti ini perlu diperbaiki karena tanah merupakan tempat berlangsungnya mikro-organisme serta aktivitas biologi lainnya. (SYAIDAH, n.d.). Banyaknya Jenis tanaman padi yang ada di indonesia membuat kami ingin meneliti seberapa baik mutu dari tanaman padi tersebut. Jenis metode data mining ada bermacam macam teknik sehingga kami ingin membuat perbandingan antara beberapa metode di antaranya metode c4.5 dan metode knn. Dari penelitian sebelumnya banyak yang menggungkapkan bahwa klasifikasi data mengunakan algoritma c4.5 sudah bagus. Hal itulah yang mendorong kami untuk mencoba komparasi antara teknik c4.5 dengan teknik knn untuk memperoleh hasil akurasi yang baik.

Tinjauan Pustaka

Bondowoso adalah salah satu dari sebagian kabupaten di Jawa Timur yang juga membudidayakan padi organik dengan tujuan meningkatkan pendapatan kabupaten Bondowoso (Faid, 2017). Selain itu tanaman padi organik juga mengandung gizi yang terbilang cukup tinggi untuk di konsumsi. Sebelumnya hasil dari padi organik ini hanya bisa di nikmati oleh kalangan masyarakat menegah ke atas, Tapi saat ini tidak menuntut kemungkinan masyarakat menegah kebawah juga bisa merasakan manfaat dari padi organik ini. (Faid, 2017) Apabila mutu padi organik diketahui maka pihak dari dinas pertanian Bondowoso bisa menentukan harga yang sesuai dengan kualitas dan kepercayaan mitra yang berkerja sama dengan pihak dinas pertanian. Maka dari itu perlu dilakukannya sebuah penelitian dalam mencari pola dari mutu padi organik dengan menggunakan metode C4.5 dan metode KNN (K-Nearest Neighbor).

Manfaat Penelitian

Manfaat dari penelitian ini adalah untuk menguji keakuratan analisa pada mutu padi organik dengan menggunakan kedua metode yaitu metode C4.5

dan metode KNN. Data yang akan di analisa adalah data mutu padi organik dari dinas pertanian bondowoso melalui situs kaggle.com (Arie, 2019).

Metode

Pada bab ini akan membahas tentang jenis penelitian, sumber data, dan metode yang di usulkan.

1. Jenis Penelitian

Penelitian ini termasuk jenis *mixed method* atau campuran dari kualitatif dan kuantitatif dimana akan menghasilkan suatu akurasi yang bisa di jadikan pembanding antara kedua metode.

2. Sumber Data

Sumber data pada penelitian kali ini yaitu menggunakan data yang diperoleh dari Dinas Pertanian Bondowoso tahun 2017 dan dapat di akses melalui kaggle.com. Data tersebut kemudian diolah sehingga menjadi dataset yang dapat digunakan dalam penelitian. Jenis data yang digunakan adalah data yang berekstensi CSV agar terbaca oleh sistem mining pada python.(Faid, 2017)

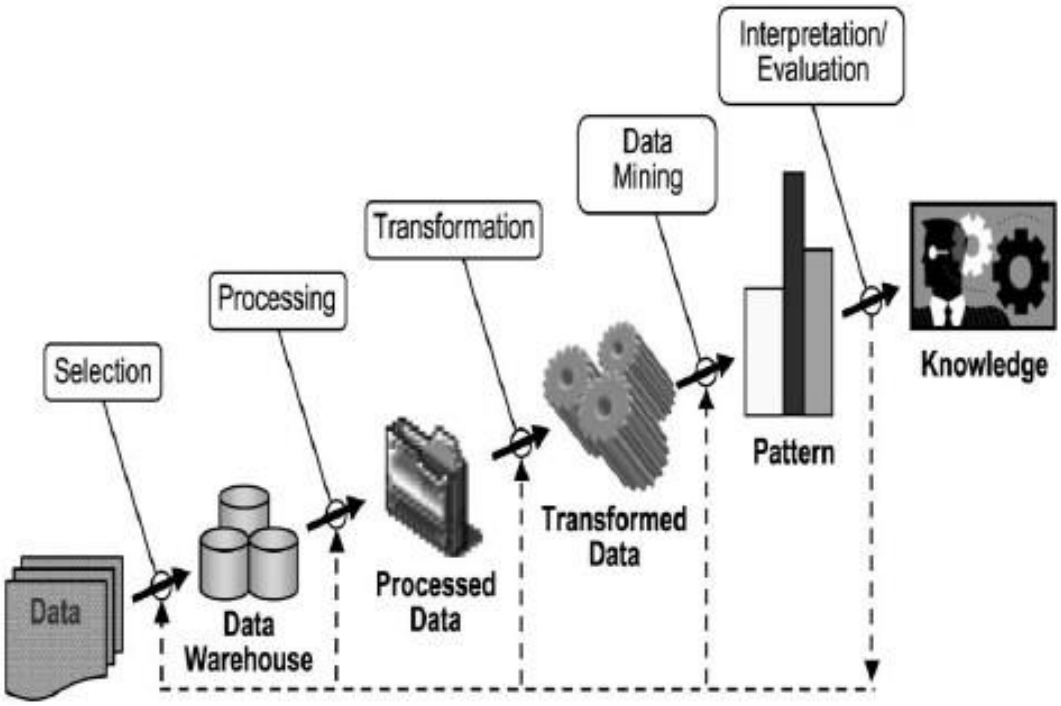
3. Metode yang diusulkan

Adapun metode yang dimaksud yaitu menggunakan k-nearest neighbor (k-NN) dan C4.5 (Pohon keputusan atau Decision Tree). Dimana akan dibandingkan hasil klasifikasi dengan Grade antar objek berdasarkan data yang jaraknya paling dekat dengan objek tersebut. (Rachman et al., n.d.-b) . Serangkaian proses tahapan data mining tersebut memiliki tahap sebagai berikut (Tan, 2004):

1. Pembersihan data (untuk membuang data yang tidak konsisten dan noise)
2. Integrasi data (penggabungan data dari beberapa sumber)
3. Transformasi data (data diubah menjadi bentuk yang sesuai untuk di-mining)

4. Aplikasi teknik Data Mining, proses ekstraksi pola dari data yang ada
5. Evaluasi pola yang ditemukan (proses interprestasi pola menjadi pengetahuan yang dapat digunakan untuk mendukung pengambilan keputusan)
6. Presentasi pengetahuan (dengan teknik visualisasi)

Tahap ini merupakan bagian dari proses pencarian pengetahuan yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Langkah terakir KDD adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami pengguna.



Tahapan Data Mining

1. Algoritma KNN

KNN merupakan algoritma klasifikasi yang paling sederhana dalam mengklasifikasikan sebuah gambar kedalam sebuah label. Metode ini mudah dipahami dibandingkan metode lain karena mengklasifikasikan berdasarkan jarak terdekat dengan objek lain (tetangga). Termasuk dalam supervised

learning, dimana hasil query instance yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam K-NN. Tujuan dari algoritma ini adalah untuk mengklasifikasikan obyek baru berdasarkan atribut dan sample-sample dari training data. AlgoritmaKNN didasarkan perbandingan contoh tes yang diberikan contoh pelatihan yang mirip. Ide dari metode KNN ini adalah untuk mengidentifikasi k sampel dalam training set yang independen variable x mirip dengan y, dan menggunakan sample k ini untuk mengklasifikasi sample baru ini kedalam kelas. Teknik pencarian tetangga terdekat rule umum dilakukan dengan menggunakan formula jarak geometrician. Berikut beberapa formula rule digunakan dalam algoritma knn.

Langkah - langkah menghitung algoritma knn

Secara umum, cara kerja algoritma KNN adalah sebagai berikut. Tentukan jumlah tetangga (K) yang akan digunakan untuk pertimbangan penentuan kelas. Hitung jarak dari data baru ke masing-masing data point di dataset. Ambil sejumlah K data dengan jarak terdekat, kemudian tentukan kelas dari data baru tersebut.

Euclidean Distance Jarak geometrician adalah formula untuk mencari jarak antara a pair of titik dalam ruang dua dimensi.

$$D(xi,xj) = \sqrt{\sum_{i=1}^{n} \left(a_r(x_i) - a_r(x_j)\right)^2}$$

Keterangan :

$D(x_i,x_j)$: Jarak Eucledean (Eucledean Distance)

(x_i) : Record ke-i

(x_j) : Record ke-j

a_r = Data ke-r

i_j = 1,2,3,..n

Hamming Distance adalah cara mencari jarak antar a pair of titik rule dihitung dengan panjang vektor biner rule dibentuk oleh dua titik tersebut dalam block kode biner. Manhattan Distance atau auto Geometri adalah formula untuk mencari jarak d antar a pair of vektor p, q pada ruang dimensi n. Minkowski distance adalah formula pengukuran antar a pair of titik pada ruang vektor traditional rule merupakan hibridisasi rule menjeneralisasi geometrician distance dan mahattan distance.

2. Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (Decision Tree). Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Pohon keputusan berguna untuk mengekspolari data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target.Algortima C4.5 didasarkan pada analisa grade pada data mutu padi organik dengan membandingkan grade pada data yaitu antara Grade A, Grade B, Grade C dan Grade D. Dimana yang terbaik adalah grade A. Sehingga akan diperoleh prediksi akurasi yang presisi. Membentuk suatu pohon keputusan yang menderkripsikan atribut pada setiap cabangnya. Kemudian cabang dari artibut tadi akan diuji dan akan mengambarkan kelas (Sularno & Anggraini, 2017).

Langkah untuk menghitung algortima C4.5

1. Menyiapkan data latih. Data yang pernah terjadi sebelumnya dan telah dikelompokkan ke dalam kelas – kelas tertentu.
2. Menentukan akar pohon. Pengambilan akat dipilih melalui atribut terpilih, dengan cara menghitung nilai gain dari masing – masing atribut. Kemudian nilai gain yang paling tinggi akan menjadi akar pertama. Sebelum nilai gain dihitung kita harus menghitung nilai entropy terlebih dahulu (Suhartini, 2019).

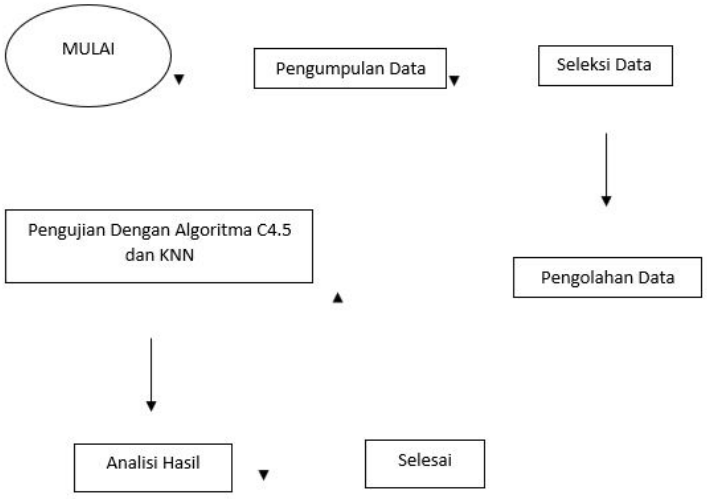
$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

Keterangan :

S : himpunan kasus
A : Atribut
n : jumlah partisi atribut A
|Si| : jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

3. Menghitung nilai gain ratio.
4. Nilai Entropy(S) adalah jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sample S. Entropy bisa dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. Semakin kecil nilai Entropy maka semakin baik untuk digunakan dalam mengekstraksi suatu kelas.
5. Mengulangi langkah ke 2 hingga semua record terpartisi
6. Selanjutnya akan ada proses partisi pohon keputusan. Partisi pohon keputusan akan berhenti jika semua record dalam simpul N mendapatkan kelas yang sama.

3. Gambaran Umum Alur Penelitian



P-ISSN: 2962-3227 | E-ISSN: 2962-3235
Prosiding Seminar Hi-Tech (Vol. 1, No. 1, 2022)
https://ejournal.unuja.ac.id/index.php/hitech

Hasil dan Pembahasan

Pengumpulan data primer yang digunakan pada penelitian kali ini yaitu bersumber dataset yang di peroleh dari Dinas Pertanian Bodowoso pada tahun 2017 dengan judul Mutu Padi Organik. Dataset adalah istilah informal yang merujuk pada kumpulan data . Secara umum, dataset berisi lebih dari satu variabel dan menyangkut suatu topik tertentu. Dataset digunakan untuk klasifikasi dengan metode data mining. Pada dataset tersebut masih terdapat string sehingga perlu dilakukannya inisiasi atau normalisasi data dari string menjadi numerik agar proses mining dengan metode KNN dapat berjalan. Kemudian pada metode C4.5 tidak perlu di melakukan inisiasi. Pada Tabel 1 menjukkan data mentah yang akan di konvert menjadi ekstensi (.csv) agar dapat dilakukannya mining (Nasution et al., 2019; Nuari et al., 2018).

No.	Variates	P.	B.	Warna	rasa	teknik	Musim	Penyakit	PH	GradeMutu
1	panda wangi	6.3	1.2	putih	sangat pulen	Jajar Lego	Hujan	Burung	2	Grade D
2	ciheran	7.2	2.3	putih	pulen	Jajar Lego	Hujan	penggerek batang	2	Grade C
3	mi kongga	6.1	3.3	putih	sangat pulen	SRI	Hujan	wereng coklat	2	Grade D
4	IR 64	6.1	4.2	putih	pulen	SRI	Hujan	wereng hijau	2	Grade B
5	Beras Merah	7.3	5.1	merah	sangat pulen	Jajar Lego	Hujan	tikus	2	Grade C
6	Beras Hitam	7.2	6.3	hitam	pulen	Jajar Lego	Hujan	Burung	2	Grade C
7	panda wangi	6.3	1.2	coklat	sangat pulen	SRI	Kemarau	penggerek batang	3	Grade B
8	ciheran	7.2	2.3	putih	pulen	SRI	Kemarau	wereng coklat	3	Grade C
9	mi kongga	6.1	3.3	putih	sangat pulen	Jajar Lego	Kemarau	wereng hijau	3	Grade B
10	IR 64	6.1	4.2	putih	pulen	Jajar Lego	Kemarau	tikus	3	Grade D
11	Beras Merah	7.3	5.1	merah	sangat pulen	SRI	Kemarau	Burung	3	Grade C
12	Beras Hitam	7.2	6.3	hitam	pulen	SRI	Kemarau	penggerek batang	3	Grade C

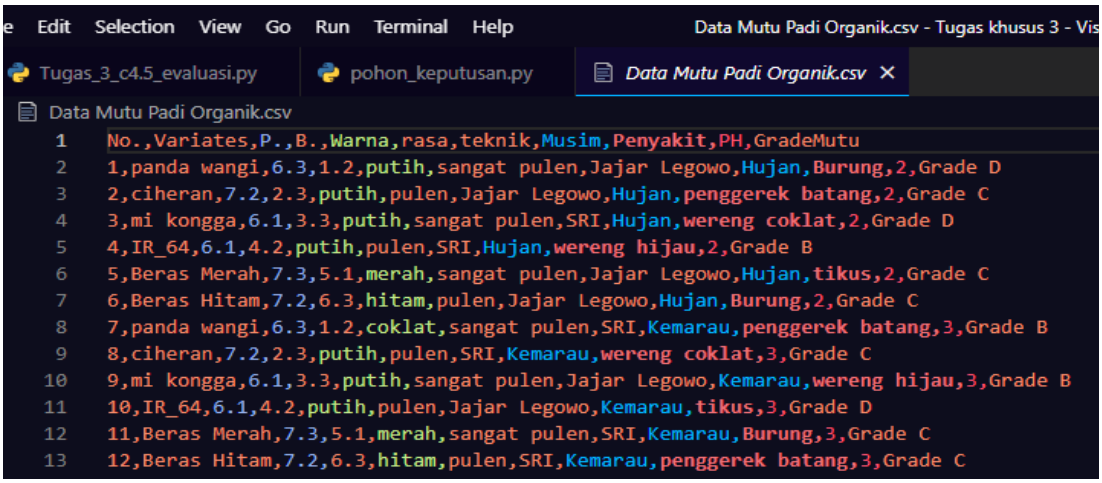
Gambar 1 sebelum dilakukakan inisiasi untuk algortima C4.5 dan berkestensi .xlsx

Pada Gambar 1 tidak dilakukan inisiasi karena label yang digunakan berbentuk string

No.	Variates	P.	B.	Warna	rasa	teknik	Musim	Penyakit	PH	Grade Mutu
1	panda wangi	6.3	1.2	putih	sangat pulen	Jajar Legowo	Hujan	Burung	2	4
2	ciheran	7.2	2.3	putih	pulen	Jajar Legowo	Hujan	penggerek batang	2	3
3	mi kongga	6.1	3.3	putih	sangat pulen	SRI	Hujan	wereng coklat	2	4
4	IR 64	6.1	4.2	putih	pulen	SRI	Hujan	wereng hijau	2	2
5	Beras Merah	7.3	5.1	merah	sangat pulen	Jajar Legowo	Hujan	tikus	2	3
6	Beras Hitam	7.2	6.3	hitam	pulen	Jajar Legowo	Hujan	Burung	2	3
7	panda wangi	6.3	1.2	coklat	sangat pulen	SRI	Kemarau	penggerek batang	3	2
8	ciheran	7.2	2.3	putih	pulen	SRI	Kemarau	wereng coklat	3	3
9	mi kongga	6.1	3.3	putih	sangat pulen	Jajar Legowo	Kemarau	wereng hijau	3	2
10	IR 64	6.1	4.2	putih	pulen	Jajar Legowo	Kemarau	tikus	3	4
11	Beras Merah	7.3	5.1	merah	sangat pulen	SRI	Kemarau	Burung	3	3
12	Beras Hitam	7.2	6.3	hitam	pulen	SRI	Kemarau	penggerek batang	3	3

Gambar 2 setelah di inisiasi untuk algortima KNN .xlsx

Pada gambar 2 dilakukan inisiasi yaitu Grade A = 1, Grade B = 2, Grade C = 3 dan Grade D = 4. Maka dengan dilakukannya inisiasi algortima KNN akan berjalan dengan baik.



Gambar 3 Data file.csv dengan vscode

Gambar 3 adalah bentuk file csv jika file tersebut di buka dengan vscode format ini yang akan terbaca karena sudah berbentuk file berekstensi (.csv).

Penelitian ini bertujuan untuk menguji keakuratan analisa mutu padi organik menggunakan komparasi algoritma C4.5 dan algoritma KNN, dimana information principle dianalisa yaitu grade pada mutu padi organik. Dataset di peroleh Dari Persian dinas pertanian bondowoso melalui kaggle. com (Rachman et al., n.d.-a; Sularno & Anggraini, 2017).

Hasil Analisa

Setelah dilakukan uji coba menggunakan program python dengan memanfaatkan library numpy, matplotlib, pandas, clasification report, dan acurasy_score didapatkan hasil akurasi dari dua algoritma tersebut dan ternyata algoritma C4.5 lebih bagus akuarasinya di bandingkan algortima KNN seperti gambar di bawah ini.

Untuk source code KNN-nya bisa di download di link ini *https://bit.ly/3axO1YT*

P-ISSN: 2962-3227 | E-ISSN: 2962-3235
Prosiding Seminar Hi-Tech (Vol. 1, No. 1, 2022)
https://ejournal.unuja.ac.id/index.php/hitech

5. Dataset Teratas

Pada langkah ini kita dapat melihat bahwa untuk memunculkan dataset teratas pada data yang berbentuk csv menggunakan library pada python yaitu pandas. Gambar di bawah ini menunjukkan tampilan lima dataset teratas pada data mutu padi organik. Disini ada informasi kolom varietas, p dan b, warna, rasa, teknik penanaman, musim, penyakit, ph dan grade mutu.

	No.	Variates	P.	B.	Warna	rasa	teknik	Musim	Penyakit	PH	Grade Mutu
0	1	panda wangi	6.3	1.2	putih	sangat pulen	Jajar Legowo	Hujan	Burung	2	4
1	2	ciheran	7.2	2.3	putih	pulen	Jajar Legowo	Hujan	penggerek batang	2	3
2	3	mi kongga	6.1	3.3	putih	sangat pulen	SRI	Hujan	wereng coklat	2	4
3	4	IR 64	6.1	4.2	putih	pulen	SRI	Hujan	wereng hijau	2	2
4	5	Beras Merah	7.3	5.1	merah	sangat pulen	Jajar Legowo	Hujan	tikus	2	3

Dataset teratas

Data Training dan Data Tetsting

Setelah itu kita bagi antara data training dan dan data testing. Pembagian dilakukan dengan memecah data set menjadi data testing dan data training. Data pada algoritma ini umumnya dibagi menjadi 2 bagian, yaitu data training dan data testing. Data training nantinya akan digunakan untuk melatih algoritma dalam mencari model yang sesuai, sedangkan data testing akan dipakai untuk menguji dan mengetahui performa model yang didapatkan pada tahapan testing.

Data pada algoritma ini umumnya dibagi menjadi 2 bagian, yaitu data training dan data testing. Data training nantinya akan digunakan untuk melatih algoritma dalam mencari model yang sesuai, sedangkan data testing akan dipakai untuk menguji dan mengetahui performa model yang didapatkan pada tahapan testing.

Membuat kodingan untuk data training dan data testing menggunakan library sklearn pada python.

```
[20] from sklearn.preprocessing import StandardScaler
     sc_X = StandardScaler()
     x_train = sc_X.fit_transform(x_train)
     x_test = sc_X.transform(x_test)
```

Kodingan untuk membuat data training dan data tersting

Berikutnya adalah menampilkan data training dan data testing

```
print(x_train)

[[ 0.77736494 -0.85189297]
 [-0.79568329  0.29441712]
 [ 0.97399597  0.92488768]
 ...
 [-1.18894534  0.29441712]
 [-0.79568329  0.29441712]
 [ 1.36725803 -0.27873792]]
```

Data training

```
[22] print(x_test)

[[ 0.58073391 -0.22142242]
 [ 0.97399597  0.81025667]
 [-0.59905226  0.35173263]
 ...
 [-1.18894534 -0.22142242]
 [ 0.97399597 -0.73726196]
 [ 1.36725803 -1.42504802]]
```

Data testing

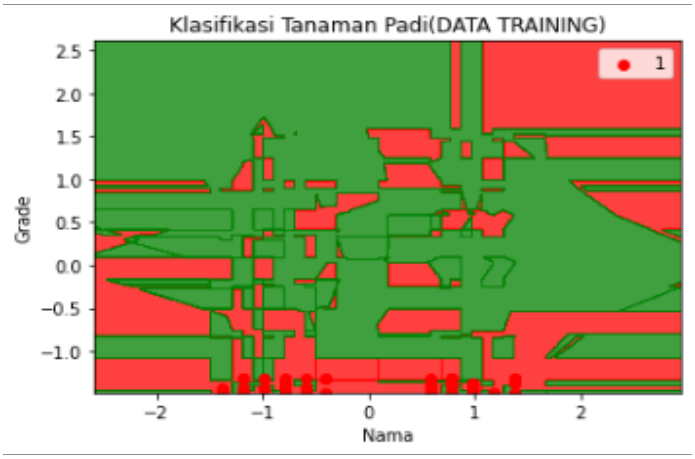
Data Training dan testing telah selesai di buat selanjutnya yaitu Akurasi Algoritma KNN. Dengan menggunakan algoritma ini dan data yang di analisa adalah grade mutu pada padi yang datanya di lakukan normalisasi data dari string menjadi integer sehingga proses dari algoritma KNN dapat berjalan. Dengan begitu maka tingkat ke Akurasian dari algoritma ini yaitu mencapai 40 % seperti pada gambar di bawah ini (Faid, 2017; Nasution et al., 2019).

P-ISSN: 2962-3227 | E-ISSN: 2962-3235
Prosiding Seminar Hi-Tech (Vol. 1, No. 1, 2022)
https://ejournal.unuja.ac.id/index.php/hitech

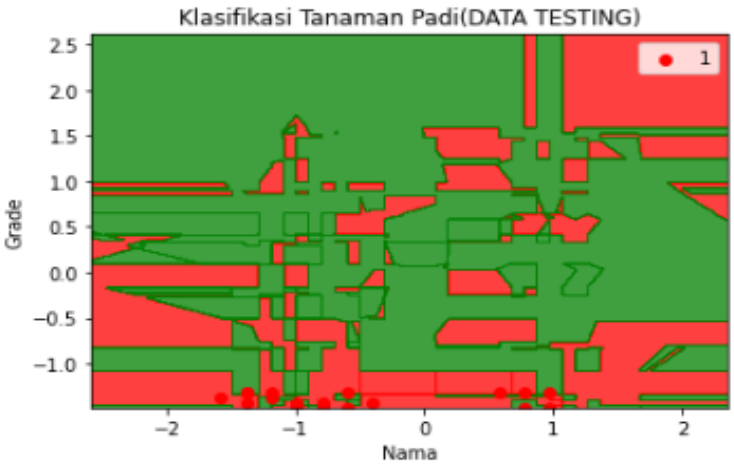
```
ini tingkat akurasi algoritma K-NN
Akurasi :               precision    recall  f1-score   support

           1       0.05      0.05      0.05        19
           2       0.34      0.34      0.34       402
           3       0.48      0.54      0.51       576
           4       0.36      0.24      0.29       241

    accuracy                           0.41      1238
   macro avg       0.31      0.29      0.30      1238
weighted avg       0.40      0.41      0.40      1238

Tingkat Akurasi :40 persen
```

Gambar Akurasi algoritma KNN

Didapatkan hasil roundmap pada algortima knn yang di peroleh dari data training dan data testing, seperti pada gambar di bawah ini.



Hasil Data Training Algoritma KNN



Hasil DataTesting Algortima KNN

Pada algoritma C4.5 kita menggunakan sampel Grade Mutu seperti gambar dibawah ini yang menampilkan banyak data pada dataset. Data yang akan di tampilkan berjumlah 4952. Berikut ini tampilan data berbentuk csv.

```
[ ] import numpy as np
    import matplotlib.pyplot as plt
    import pandas as pd

    #baca database
    dataset = pd.read_csv('Data Mutu Padi Organik.csv')
    dataset
```

	No.	Variates	P.	B.	Warna	rasa	teknik	Musim	Penyakit	PH	Grade Mutu
0	1	panda wangi	6.3	1.2	putih	sangat pulen	Jajar Legowo	Hujan	Burung	2	4
1	2	ciheran	7.2	2.3	putih	pulen	Jajar Legowo	Hujan	penggerek batang	2	3
2	3	mi kongga	6.1	3.3	putih	sangat pulen	SRI	Hujan	wereng coklat	2	4
3	4	IR 64	6.1	4.2	putih	pulen	SRI	Hujan	wereng hijau	2	2
4	5	Beras Merah	7.3	5.1	merah	sangat pulen	Jajar Legowo	Hujan	tikus	2	3
...	...	...	...	...	...	...	...	...	...	...	...
4947	4948	IR 64	7.2	4.2	putih	pulen	SRI	Hujan	wereng coklat	2	4
4948	4949	Beras Merah	6.1	5.3	merah	sangat pulen	Jajar Legowo	Hujan	wereng hijau	2	2
4949	4950	Beras Hitam	7.4	6.2	hitam	pulen	Jajar Legowo	Hujan	tikus	2	3
4950	4951	panda wangi	6.2	1.2	putih	sangat pulen	SRI	Kemarau	Burung	3	3
4951	4952	ciheran	6.4	2.3	coklat	pulen	SRI	Kemarau	penggerek batang	3	4

```
✓ 0 d  selesai pada 08.39
```

Dataset padi organik

Selajutnya preprocessing pada data. Preprocessing data adalah proses yang mengubah data mentah ke dalam bentuk yang lebih mudah dipahami. Proses ini penting dilakukan karena data mentah sering kali tidak memiliki format yang teratur.

Langkah-langkah data preprocessing

1. Pembersihan data.

Sebagai langkah awal, yang dilakukan adalah pembersihan data terlebih dahulu. Maksudnya di sini adalah menyeleksi data mentahan yang diperoleh. Dari proses seleksi inilah dapat dilakukan pemilahan data, apakah harus dihapus atau tidak. Dengan cara ini, bisa menghindari kesalahpahaman saat melakukan analisis data.

2. Penggabungan data.

Selanjutnya, yaitu melakukan integrasi atau menggabungkan sejumlah data di sebuah data set. Untuk menggabungkan data ini, harus melihat

P-ISSN: 2962-3227 | E-ISSN: 2962-3235
Prosiding Seminar Hi-Tech (Vol. 1, No. 1, 2022)
https://ejournal.unuja.ac.id/index.php/hitech

kembali sumber-sumber data yang diperoleh. Hal itu penting dilakukan agar data yang akan digabungkan memiliki format sama.

3. Pengubahan bentuk data.

Langkah data preprocessing yang ketiga adalah transformasi data atau pengubahan bentuk data yang ada. Ingat, data yang dikumpulkan dari banyak sumber kemungkinan besar terdapat perbedaan format. Maka dari itu, harus mengubah bentuk data ini agar proses analisis datanya menjadi lebih mudah.

4. Pengurangan data.

Terakhir yang harus di lakukan dalam langkah data preprocessing adalah mengurangi data atau yang biasa dikenal dengan data reduction. Mengurangi data di sini maksudnya adalah mengurangi sampel yang diambil. Meski demikian, pengurangan data ini tidak boleh mengubah hasil dari analisis data.

Setelah itu dilakukan preprocessing pada data seperti pada gambar di bawah ini.

```
[ ] from sklearn.preprocessing import LabelEncoder

[ ] enc = LabelEncoder()

[ ] dataset['Variates'] = enc.fit_transform(dataset['Variates'].values)
    dataset['Warna'] = enc.fit_transform(dataset['Warna'].values)
    dataset['rasa'] = enc.fit_transform(dataset['rasa'].values)
    dataset['teknik'] = enc.fit_transform(dataset['teknik'].values)
    dataset['Musim'] = enc.fit_transform(dataset['Musim'].values)
    dataset['Penyakit'] = enc.fit_transform(dataset['Penyakit'].values)
    dataset['Grade Mutu'] = enc.fit_transform(dataset['Grade Mutu'].values)

[ ] dataset
```

	No.	Variates	P.	B.	Warna	rasa	teknik	Musim	Penyakit	PH	Grade Mutu
0	1	5	6.3	1.2	3	1	0	0	0	2	3
1	2	3	7.2	2.3	3	0	0	0	1	2	2
2	3	4	6.1	3.3	3	1	1	0	3	2	3
3	4	2	6.1	4.2	3	0	1	0	4	2	1
4	5	1	7.3	5.1	2	1	0	0	2	2	2
...	...	...	...	...	...	...	...	...	...	...	...
4947	4948	2	7.2	4.2	3	0	1	0	3	2	3
4948	4949	1	6.1	5.3	2	1	0	0	4	2	1
4949	4950	0	7.4	6.2	1	0	0	0	2	2	2
4950	4951	5	6.2	1.2	3	1	1	1	0	3	2
4951	4952	3	6.4	2.3	0	0	1	1	1	3	3

4952 rows × 11 columns

Preprocessing

Langkah selanjutnya yaitu memisahkan data mutu padi dengan data yang lain seperti gambar berikut.

```
[ ]  atr_dataset = dataset.drop(columns='Grade Mutu')
     atr_dataset
```

	No.	Variates	P.	B.	Warna	rasa	teknik	Musim	Penyakit	PH
0	1	5	6.3	1.2	3	1	0	0	0	2
1	2	3	7.2	2.3	3	0	0	0	1	2
2	3	4	6.1	3.3	3	1	1	0	3	2
3	4	2	6.1	4.2	3	0	1	0	4	2
4	5	1	7.3	5.1	2	1	0	0	2	2
...	...	...	...	...	...	...	...	...	...	...
4947	4948	2	7.2	4.2	3	0	1	0	3	2
4948	4949	1	6.1	5.3	2	1	0	0	4	2
4949	4950	0	7.4	6.2	1	0	0	0	2	2
4950	4951	5	6.2	1.2	3	1	1	1	0	3
4951	4952	3	6.4	2.3	0	0	1	1	1	3

4952 rows × 10 columns

Data set drop

Dari hasil drop data pada gambar di atas di peroleh hasil sedemikian rupa. Data yang di gunakan hanya pada Grade Mutu saja.

```
[ ]  cls_dataset = dataset['Grade Mutu']
     cls_dataset
```

```
0       3
1       2
2       3
3       1
4       2
       ..
4947    3
4948    1
4949    2
4950    2
4951    3
Name: Grade Mutu, Length: 4952, dtype: int64
```

Sampel data grade mutu

Membuat model DecisionTreeClassfier/ model klasifikasi algoritma C4.5

P-ISSN: 2962-3227 | E-ISSN: 2962-3235
Prosiding Seminar Hi-Tech (Vol. 1, No. 1, 2022)
https://ejournal.unuja.ac.id/index.php/hitech

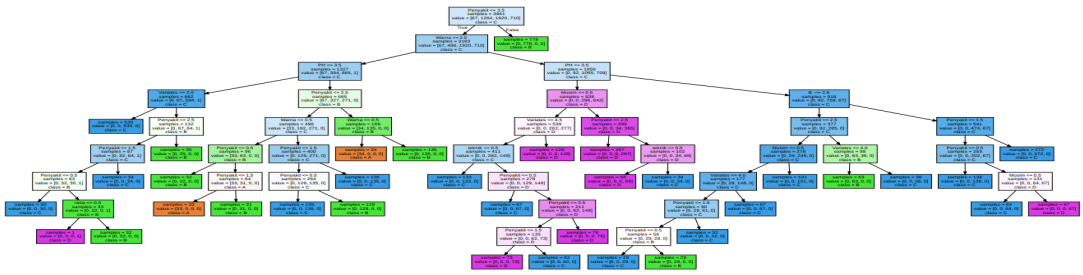
```
[ ] from sklearn.model_selection import train_test_split
    from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
    from sklearn.tree import DecisionTreeClassifier

[ ] x_train, x_test, y_train, y_test = train_test_split(atr_dataset, cls_dataset, test_size=0.2, random_state=1)
    tree_dataset = DecisionTreeClassifier(random_state=1)
    tree_dataset.fit(x_train, y_train)

    DecisionTreeClassifier(random_state=1)
```

Kemudian dilakukanlah perhitungan akurasi pada sampel tersebut sehingga menghasilkan akurasi 99%. Seperti pada gambar di bawah ini. Source code bisa di download *https://bit.ly/3uDD8LI*

```
Confusion Matrik
Tingkat Akurasi Algoritma C4.5
Akurasi :                precision    recall  f1-score   support

           0       1.00      1.00      1.00        15
           1       1.00      1.00      1.00       345
           2       1.00      1.00      1.00       472
           3       0.99      1.00      1.00       159

    accuracy                           1.00       991
   macro avg       1.00      1.00      1.00       991
weighted avg       1.00      1.00      1.00       991

Tingkat Akurasi : 99 persen
```

Gambar Akurasi Algoritma C4.5

Pohon keputusan atau decision tree merupakan teknik data mining yang digunakan untuk mengeksplorasi data dengan membagi kumpulan data yang besar menjadi himpunan record yang lebih kecil dan memperhatikan variabel tujuannya.



Kesimpulan

Metode C4.5 dan Metode KNN dapat digunakan untuk melakukan klasifikasi pada tanaman padi. Adapun penerapan algoritma masih cukup sederhana dan untuk mendapatkan nilai akurasi dan yang baik, perlu

dilakukan pengujian untuk nilai-K pada metode KNN dan pada metode C4.5 perlu dilakukan pengujian pula pada rule nya yang akan digunakan serta dilakukan perbandingan dengan metode lainnya sehingga akan didapat model terbaik. Dari percobaan di atas dapat di simpulkan bahwa klasifikasi mengunakan metode algoritma C4.5 menghasilkan akurasi yang baik di bandingkan dengan metode knn.(Rachman et al., n.d.-a; Suhartini, 2019).

Implikasi

Berdasarkan hasil penelitian tersebut dapat dikemukakan implikasi secara teoritis dan praktis sebagai berikut:

Implikasi Teoritis

Pemilihan metode KNN pada klasifikasi data mutu padi organik ini berpengaruh terhadap mutu dan kualitas suatu jenis padi organik. Metode C4.5 pada klasifikasi tanaman padi sangat berpengaruh untuk mendapatkan suatu pohon keputusan. Kedua algoritma ini bertujuan untuk menemukan hasil presisi dan akurasi yang akan ditemukan pada data mutu padi organik.

Implikasi Praktis

Hasil penelitian yang di gunakan mengatakan bahwa hasil dari kedua metode algortima tersebut sangatlah berbeda, pada algortima KNN mempunyai akurasi 40 % sedangkan pada algortima C4.5 mendapatkan akurasi 99%.

Rekomendasi

Dari hasil analisis dan kesimpulan, peneliti memberikan beberapa rekomendasi diantaranya klasifikasi menggunakan metode algoritma c4.5 lebih bagus dari pada algortima knn sehingga pada kasus yang lain kami sangat merekomendasi algortima tersebut.

Daftar Pustaka

Adhy, D. R. (2021). Rancang Bangun Sistem Prediksi Varietas Padi Yang Cocok Dengan Lahan Menggunakan Metode Data Mining Algoritma C4. 5. *Jurnal Ilmiah Sains, Teknologi Dan Rekayasa*, *1*(1), 32–39.

Arie, J. S. (2019). Implementasi Algoritma KNN Dalam Memprediksi Curah Hujan dan Temperatur Untuk Tanaman Padi. *SISITI: Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, *8*(1).

Faid, M. (2017). Klasifikasi mutu padi organik menggunakan C4. 5 di dinas pertanian bondowoso. *Prosiding SNATIF*, 155–162.

Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *CESS (Journal of Computer Engineering, System and Science)*, *4*(1), 78–82.

Nuari, R., Apriliyani, A., Juwari, J., & Kusrini, K. (2018). IMPLEMENTASI METODE K-NEAREST NEIGHBOR (KNN) UNTUK MEMPREDIKSI VARIETAS PADI YANG COCOK UNTUK LAHAN PERTANIAN. *Jurnal Informa: Jurnal Penelitian Dan Pengabdian Masyarakat*, *4*(2), 28–34.

Rachman, A., Furqon, M. T., & Ramdani, F. (n.d.-a). Klasifikasi Varietas Unggul Padi menggunakan Algoritme C4. 5. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer E-ISSN*, *2548*, 964X.

Rachman, A., Furqon, M. T., & Ramdani, F. (n.d.-b). Klasifikasi Varietas Unggul Padi menggunakan Algoritme C4. 5. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer E-ISSN*, *2548*, 964X.

Sari, S. M., & Hasibuan, N. A. (2019). ANALISA DATA PERTANIAN TANAMAN PANGAN UNTUK MEMPREDIKSI HASIL PANEN DENGAN DATA MINING ALGORITMA C. 45 (STUDI KASUS: DINAS TANAMAN PANGAN dan HOLTIKUTURA PROVINSI SUMUT). *Pelita Informatika: Informasi Dan Informatika*, *7*(4), 473–480.

Suhartini, S. (2019). Klasifikasi Pengaruh Faktor Cuaca Terhadap Hasil Produksi Tanaman Pangan Di Yogyakarta Menggunakan Metode Decision Tree. *Naskah Publikasi Program Studi Sistem Informasi*.

Sularno, S., & Anggraini, P. (2017). Penerapan Algoritma C4. 5 Untuk Klasifikasi Tingkat Keganasan Hama Pada Tanaman Padi (Studi Kasus: Dinas Pertanian Kabupaten Kerinci). *Jurnal Sains Dan Informatika: Research of Science and Informatic*, *3*(2), 161–170.

SYAIDAH, S. S. N. U. R. (n.d.). *KLASIFIKASI KUALITAS PADI ORGANIK DENGAN MENGGUNAKAN ALGORITMA C4. 5 DI DINAS KETAHANAN PANGAN, PERTANIAN DAN PERIKANAN*.