

Klasifikasi Varietas Unggul Padi menggunakan Algoritme C4.5

Arinda Rachman¹, Muhammad Tanzil Furqon², Fatwa Ramdani³

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: ¹rachmanarinda@gmail.com, ²m.tanzil.furqon@ub.ac.id, ³fatwaramdani@ub.ac.id

Abstrak

Varietas unggul merupakan komponen pendukung yang penting dalam peningkatan produksi padi nasional. Setiap jenis varietas unggul memiliki ciri yang berbeda. petani membutuhkan pengetahuan untuk dapat memilih jenis padi sesuai yang diinginkan sekaligus menjaga agar produksi padi tetap stabil. Untuk mengatasi masalah dalam penentuan benih padi, salah satu solusinya adalah dengan menggunakan klasifikasi *data mining*. klasifikasi menggunakan metode *decision tree*, dengan algoritme C4.5. Algoritme C4.5 adalah algoritme yang membangun pohon keputusan dari data dengan pola rekursif. C4.5 mendatangi setiap simpul dari pohon keputusan, kemudian memilih cabang yang optimal sampai tidak ada lagi cabang yang dihasilkan. algoritme tersebut dapat menghasilkan klasifikasi varietas unggul padi yang diperlukan untuk ditanam di sawah. Penelitian ini memiliki data masukan berupa daun bendera, bentuk gabah, warna gabah, kerabahan dan tekstur nasi. Data masukan tersebut kemudian diolah oleh algoritme C4.5 sehingga menghasilkan luaran yang menentukan varietas unggul padi. Hasil penelitian ini menunjukkan bahwa performa klasifikasi varietas unggul padi lebih banyak dipengaruhi oleh atribut bentuk gabah dibandingkan dengan atribut lainnya. Sementara, akurasi algoritme C4.5 tertinggi pada perhitungan k-fold 1 sebanyak 28,571%. Sementara, k-fold 2 sebanyak 17,857%, k-fold 3 10,344% dan k-fold 4 sebanyak 10,714%.

Kata kunci: *varietas unggul, klasifikasi, algoritme C4.5*

Abstract

Superior varieties are important supporting component in increasing national rice production. Each type of superior variety has different characteristics. Farmers need knowledge to be able to choose the type of rice they want while keeping rice production stable. To overcome the problem in determining rice seeds, one solution is to use data mining classification. The classification are using the decision tree method, with the C4.5 algorithm. C4.5 algorithm is an algorithm that builds decision trees from data with recursive patterns. C4.5 goes to each node of the decision tree, then selects the optimal branch until no more branches are generated. the algorithm can produce a classification of superior rice varieties needed to be planted in the fields. This research has input data in the form of flag leaves, grain shape, grain color, rice texture and texture. The input data is then processed by the C4.5 algorithm to produce outcomes that determine superior varieties of rice. The results of this study indicate that the classification performance of superior rice varieties is more influenced by the grain form attributes compared to other attributes. Meanwhile, the accuracy of the highest C4.5 algorithm in the k-fold 1 calculation was 28.571%. Meanwhile, k-fold 2 was 17.857%, k-fold 3 was 10.344% and k-fold 4 was 10.714%.

Keywords: *superior varieties, classification, algorithm C4.5*

1. PENDAHULUAN

Pertanian mulai berkembang pesat tidak hanya di mancanegara, namun juga di Indonesia. Indonesia terkenal dengan sebutan negara agraris karena sebagian besar mata pencarian penduduknya di bidang pertanian. Salah satu

produk pertanian penghasil bahan makanan pokok utama adalah padi. Padi merupakan salah satu komoditas ekspor bagi Indonesia dan padi juga merupakan kebutuhan yang sangat penting bagi berlangsungnya kehidupan atau penghasil pangan bagi penduduk Indonesia. Seiring dengan meningkatnya jumlah penduduk serta gaya hidupnya, maka upaya dalam peningkatan

jumlah produksi padi akan semakin meningkat pula (Imanda, 2017).

Komponen pendukung yang mempunyai peran sangat penting dalam peningkatan produksi padi nasional adalah varietas unggul. Pada tahun 2017 diluncurkan jenis varietas unggul padi oleh Badan Penelitian dan Pengembangan Pertanian melalui Balai Besar Penelitian Tanaman Padi guna memberi dukungan swasembada beras. Inpa (INBRIDA PADI) dan Hipa (HIBRIDA PADI) merupakan Varietas Unggul Baru (VUB). Untuk INBRIDA Padi, dapat mewakili ekosistem padi tersebut tumbuh. Hal tersebut dapat ditunjukkan dengan penulisan nama yang menggunakan suku kata inpa. Sebagai contoh: INPARI = Inbrida Padi Sawah Irigasi, INPARA = Inbrida Padi Rawa, dan INPAGO = Inbrida Padi Gogo. Cere merupakan varietas unggul padi hasil dari persilangan tetua padi, padi cere memiliki hasil produksi yang melimpah walaupun dengan waktu tanam yang sangat singkat. Pada setiap jenis varietas unggul memiliki ciri yang berbeda, oleh karena itu agar petani dapat memilih jenis padi sesuai yang diinginkan, petani butuh pengetahuan tersebut agar produksi padi tetap terjaga kestabilannya. Untuk mendukung proses tersebut dibutuhkan proses klasifikasi varietas padi (Wahab, Satoto dan Suharna 2017).

Untuk mengatasi masalah dalam penentuan benih padi oleh petani, salah satunya dapat dilakukan dengan menggunakan metode *data mining*. *Data mining* merupakan proses dalam penguraian dan pengenalan informasi dan pengetahuan yang bermanfaat dengan bantuan teknik statistik, matematika, *artificial intiligent* dan *machine learning*. Proses *data mining* biasa digunakan untuk mengurai informasi atau pengetahuan dari *database* besar (Nofriansyah, 2012). Dalam *data mining* terdapat proses menentukan *record* dari data baru yang telah didefinisikan sebelumnya, data tersebut dimasukkan ke dalam beberapa kelas yang disebut klasifikasi (Hermawati, 2013).

Penelitian sebelumnya yang dilakukan oleh Imanda, 2017, membahas klasifikasi varietas unggul padi menggunakan algoritme *Modified K-Nearest Neighbour* (MKNN), penelitian tersebut berguna untuk meningkatkan produksi padi. Parameter yang digunakan yaitu, umur, kerontokan, kerebahan, tekstur nasi, rata-rata hasil, potensi hasil, ketahanan terhadap hama dan ketahanan terhadap penyakit. Hasil akurasi tertinggi sebanyak 79,96% dan akurasi terendah sebanyak 51,2%. Sementara dalam menentukan

klasifikasi varietas unggul padi pada penelitian ini menggunakan metode *decision tree*. Metode *decision tree* yang dipilih menggunakan algoritme C4.5. Kegunaan algoritme C4.5 adalah membangun pohon keputusan dari data dengan pola secara rekursif mendatangi setiap simpul dari keputusan kemudian memilih cabang yang optimal sampai tidak ada lagi cabang yang dihasilkan (Kholifah, 2016). Keunggulan algoritme C4.5 yaitu pohon keputusan yang dihasilkan oleh algoritme C4.5 mudah untuk ditafsirkan maknanya, mempunyai tingkat akurasi yang bisa diterima, mengatasi atribut bertipe diskret secara efisien dan dapat menangani atribut bertipe diskret dan numerik (Kamagi dan Hansun, 2014). Melalui algoritme tersebut dapat menghasilkan klasifikasi varietas unggul padi yang di perlukan untuk di tanam di tanah sawah. Penelitian yang dibuat penulis menerapkan algoritme C4.5 dengan data masukan berupa daun bendera, bentuk gabah, warna gabah, kerabahan dan tekstur nasi. Data masukan tersebut kemudian diolah menggunakan algoritme C4.5 sehingga menghasilkan keluaran yang menentukan varietas unggul padi.

Berdasarkan data yang penulis dapatkan, bisa disimpulkan bahwa petani membutuhkan sistem untuk memudahkan dalam menentukan varietas unggul padi yang akan ditanam di tanah sawah agar kedepannya bidang pertanian di Indonesia lebih maju. Implementasi sistem ini menggunakan pemrograman php berbasis web dan *database* MySQL karena kedepannya akan digunakan langsung oleh petani. Web dipilih karena pada tahun 2017 pengguna internet masih memilih web sebagai media yang mudah digunakan. Hal ini didukung dengan adanya 44,16 persen orang mengakses internet melalui *smartphone* (APJII, 2017).

2. DATA DAN METODOLOGI

Pengumpulan data pada penelitian ini diambil dari website resmi Balai Besar Penelitian Tanaman Padi Balitbangan Kementerian Pertanian (BBPADI) dan Dinas Pertanian Magetan. Penelitian ini menggunakan data padi berjumlah 113 data terdiri dari lima kelas yaitu, Inbrida Padi Sawah Irigasi (INPARI), Hibrida padi (HIPA), Inbrida Padi Gogo (INPAGO), Inbrida Padi Rawa (INPARA) dan CERE. Masing-masing kelas terdiri dari lima atribut yaitu, daun bendera, bentuk gabah, warna gabah, kerebahan dan tekstur nasi. Jumlah

data pada masing-masing kelas berbeda, pada kelas INPAGO sebanyak 11 data, CERE sebanyak 36 data, INPARI sebanyak 41 data, INPARA sebanyak 8 data dan HIPA sebanyak 7 data.

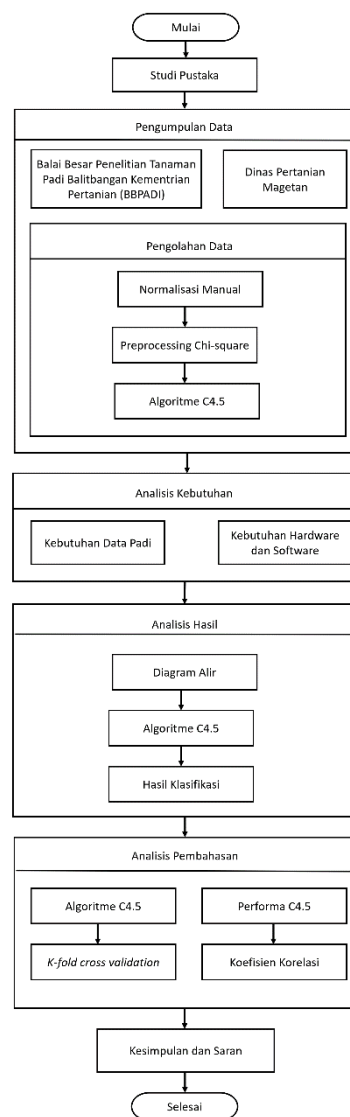
Strategi penelitian ini melakukan eksperimen membuat sistem klasifikasi padi menggunakan algoritme C4.5 dengan subyek adalah petani. Lokasi penelitian ini di Lab. Teknik dalam analisis data pembahasan menggunakan analisis kuantitatif secara statistik menggunakan uji k-fold cross validation untuk akurasi algoritme C4.5 dan Koefisien Korelasi untuk mengetahui seberapa besar performa dari klasifikasi varietas unggul padi. Tahapan yang akan dilalui dalam penyusunan penelitian ini meliputi studi pustaka, pengumpulan data, analisis kebutuhan, analisis hasil, analisis pembahasan, kesimpulan dan saran ditampilkan pada Gambar 1.

3. HASIL

3.1. Preprocessing Chi-Kuadrat

Untuk melakukan preprocessing data pada penelitian ini menggunakan metode chi-kuadrat. langkah kerjanya menyeleksi atribut mana yang memiliki nilai chi-kuadrat atau (χ^2) lebih besar daripada nilai probabilitas pada tabel persentase poin dari distribusi chi-kuadrat (p-value). Penelitian ini menggunakan degrees of freedom (df) untuk mencari nilai probabilitas nilai, kemudian pada tabel p-value memilih df 0.05 sebagai ketentuannya. Degress of freedom dipilih 0.05 karena ingin membandingkan nilai χ^2 dengan nilai p-value. Ilustrasi perhitungan yang sudah dijelaskan bisa dilihat pada Tabel 1.

Berdasarkan keterangan dari Tabel 1 atribut bentuk tanaman dan kerontokan memiliki nilai chi-kuadrat lebih kecil daripada p-value sehingga kedua atribut tersebut dihilangkan. Atribut yang digunakan dalam penelitian ini yaitu, daun bendera, bentuk gabah, warna gabah, kerebahan dan tekstur nasi. Setelah ditemukan atribut terbaik tahapan selanjutnya menghitung data menggunakan algoritme C4.5.



Gambar 1 Metodologi Penelitian

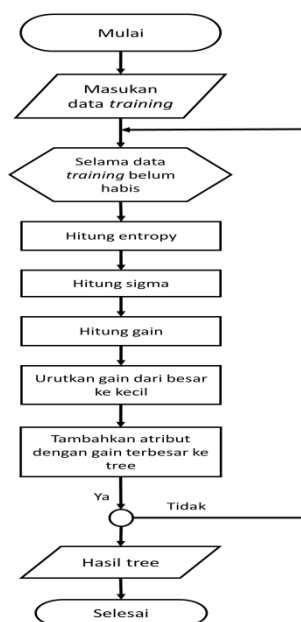
Tabel 1. Preprocessing Chi-Kuadrat

No	Nama Atribut	Jumlah χ^2	DF	P-Value
1	Bentuk tanaman	11,127	8	15.51
2	Daun bendera	81,685	56	79.08
3	Bentuk gabah	106,832	88	79.08
4	Warna gabah	93,925	32	43.77
5	Kerontokan	40,384	28	41.34
6	Kerebahan	45,526	32	43.77
7	Tekstur nasi	72,090	48	67.50

3.2. Algoritme C4.5

Digram alir algoritme C4.5 menjelaskan langkah-langkah bagaimana cara kerja pada

algoritme C4.5. Langkahnya, menyediakan data masukan yang akan digunakan sebagai data *training*, terjadi proses perulangan dengan kondisi selama data *training* belum habis maka hitung masing-masing nilai dari atribut tanaman padi dengan menghitung entropy, sigma dan gain, kemudian mengurutkan nilai gain berdasarkan besar ke kecil, setelah nilai gain terbesar didapatkan lalu tambahkan atribut tersebut ke bentuk tree, apabila keluaran sesuai maka menghasilkan tampilan tree sedangkan apabila hasil keluaran tidak sesuai maka kembali ke perhitungan perulangan kondisi kembali. Diagram alir algoritme C4.5 ditampilkan pada Gambar 2 dibawah ini:



Gambar 2. Diagram Alir Algoritme C4.5

4. PEMBAHASAN

4.1. Pengujian Akurasi C4.5

Penelitian ini menggunakan pengujian *k-fold cross validation*, yaitu dengan membagi beberapa data menjadi empat bagian uji, dari empat bagian uji tersebut masing-masing dibagi menjadi data *testing* dan data *training*. Setelah dilakukan pengujian dari masing-masing perhitungan *fold* tersebut dihasilkan nilai akurasi. Ilustrasi pengujian akurasi yang sudah dijelaskan bisa dilihat pada Tabel 2.

Tabel 2 Pengujian Akurasi C4.5

K-Fold	Data Testing	Data Training	Akurasi
1	1 - 28	29 - 113	28,571
2	29 - 56	1-28, 57 - 113	17,857
3	57 - 85	1 - 56, 86 - 113	10,344
4	86 - 113	1 - 85	10,714

Berdasarkan keterangan dari Tabel 2 hasil akurasi tertinggi pada saat perhitungan *k-fold* 1 data testing 1 - 28 dan data training 29 - 113 yaitu akurasi sebanyak 28,571%. Pada *k-fold* 2 data testing 29 - 56 dan data training 1 - 28, 57 - 113 akurasi sebanyak 17,857%. Pada pengujian *k-fold* 3 dan *k-fold* 4 terjadi hasil akurasi rendah tetapi *k-fold* 3 hasil akurasinya terendah dari *k-fold* 4, *k-fold* 3 data testing 57 - 85 dan data training 1 - 56, 86 - 113 akurasi sebanyak 10,344%, sedangkan *k-fold* 4 data testing 86 - 113 dan data training 1 - 85 akurasi sebanyak 10,714%.

Dalam akurasi pengujian *k-fold cross validation* menghasilkan akurasi kecil dikarenakan data pada atribut memiliki ciri nama yang sama dengan atribut yang lain dan kelas satu dengan kelas lainnya jumlahnya tidak sama.

4.2. Pengujian Metode Koefisien Korelasi

Untuk menjawab seberapa besar performa sistem klasifikasi varietas unggul padi menggunakan algoritme C4.5 maka dibuatlah perhitungannya dengan menggunakan metode koefisien korelasi dengan ketentuan jumlah data *testing*.

4.2.1. Koefisien Korelasi 1 - 28

Tabel 3 Koefisien Korelasi 1 -28

No	Tanaman Padi	Hasil
1	Atribut 1 (daun bendera)	0/0
2	Atribut 2 (bentuk gabah)	0,341
3	Atribut 3 (warna gabah)	0,715
4	Atribut 4 (Kerebahan)	-0,005
5	Atribut 5 (tekstur nasi)	0,144

Berdasarkan Tabel 3 koefisien korelasi data *testing* 1 - 28 menampilkan tiga hasil positif, satu hasil negatif dan satu hasil nol. Hasil positif terdapat pada atribut 2, 3, 5, hasil negatif terdapat pada atribut 4 dan hasil nol terdapat pada atribut 1. Sehingga dapat ditarik kesimpulan bahwa atribut 3 memiliki performa

paling tinggi dibandingkan atribut 1, 2, 4, 5 sedangkan pada atribut 1 memiliki performa paling rendah. Atribut 3 (warna gabah) lebih dominan daripada atribut lainnya karena warna gabah memiliki hasil koefisien korelasi sebanyak 0,715.

4.2.2. Koefisien Korelasi 29 - 56

Tabel 4 Koefisien Korelasi 29 - 56

No	Tanaman Padi	Hasil
1	Atribut 1 (daun bendera)	-0,067
2	Atribut 2 (bentuk gabah)	0,411
3	Atribut 3 (warna gabah)	0,373
4	Atribut 4 (Kerebahan)	0,039
5	Atribut 5 (tekstur nasi)	0,072

Berdasarkan Tabel 4 koefisien korelasi data *testing* 29 - 56 menampilkan empat hasil positif dan satu hasil negatif. Hasil positif terdapat pada atribut 2, 3, 4, 5 dan hasil negatif terdapat pada atribut 1. Sehingga dapat ditarik kesimpulan bahwa atribut 2 memiliki performa paling tinggi dibandingkan atribut 1, 3, 4, 5 sedangkan pada atribut 1 memiliki performa paling rendah. Atribut 2 (bentuk gabah) lebih dominan daripada atribut lainnya karena bentuk gabah memiliki hasil koefisien korelasi sebanyak 0,411.

4.2.3. Koefisien Korelasi 57 - 85

Tabel 5 Koefisien Korelasi 57 - 85

No	Tanaman Padi	Hasil
1	Atribut 1 (daun bendera)	-0,228
2	Atribut 2 (bentuk gabah)	0,332
3	Atribut 3 (warna gabah)	0,525
4	Atribut 4 (Kerebahan)	-0,069
5	Atribut 5 (tekstur nasi)	0,209

Berdasarkan Tabel 5 koefisien korelasi data *testing* 57 - 85 menampilkan tiga hasil positif dan dua hasil negatif. Hasil positif terdapat pada atribut 2, 3, 5 dan hasil negatif terdapat pada atribut 1, 4. Sehingga dapat ditarik kesimpulan bahwa atribut 3 memiliki performa paling tinggi dibandingkan atribut 1, 2, 4, 5 sedangkan pada atribut 1 memiliki performa paling rendah. Atribut 3 (warna gabah) lebih dominan daripada atribut lainnya karena warna gabah memiliki hasil koefisien korelasi sebanyak 0,525.

4.2.4. Koefisien Korelasi 86 – 113

Berdasarkan Tabel 6 koefisien korelasi data *testing* 86 - 113 menampilkan satu hasil positif dan empat hasil negatif. Hasil positif terdapat

pada atribut 2 dan hasil negatif terdapat pada atribut 1, 3, 4, 5. Sehingga dapat ditarik kesimpulan bahwa atribut 3 memiliki performa paling tinggi dibandingkan atribut 1, 3, 4, 5 sedangkan pada atribut 1 memiliki performa paling rendah. Atribut 2 (bentuk gabah) lebih dominan daripada atribut lainnya karena bentuk gabah memiliki hasil koefisien korelasi sejumlah 0,091.

Tabel 6 Koefisien Korelasi 86 - 113

No	Tanaman Padi	Hasil
1	Atribut 1 (daun bendera)	-0,326
2	Atribut 2 (bentuk gabah)	0,091
3	Atribut 3 (warna gabah)	-0,065
4	Atribut 4 (Kerebahan)	-0,258
5	Atribut 5 (tekstur nasi)	-0,291

Dalam pengujian metode koefisien korelasi dari masing-masing perhitungan empat data *testing* dapat disimpulkan bahwa performa atribut ke 2 yaitu bentuk gabah terjadi peningkatan secara positif pada setiap perhitungan data *testing*. Sementara, atribut ke 3 yaitu warna gabah terjadi peningkatan dan penurunan performa secara bergantian, atribut ke 4 yaitu kerebahan terjadi dua kali peningkatan dan dua kali penurunan performa, atribut ke 1 yaitu daun bendera terjadi penurunan performa secara negatif secara terus-menerus. Sehingga dapat disimpulkan atribut 2 bentuk gabah adalah atribut yang lebih dominan dari pada atribut lainnya. Hal tersebut juga dibuktikan dari banyaknya ciri-ciri data pada atribut bentuk gabah pada data tanaman padi.

5. KESIMPULAN DAN SARAN

Kesimpulan dalam penelitian ini, performa yang dihasilkan menggunakan metode koefisien korelasi menyatakan bahwa atribut bentuk gabah memiliki performa yang lebih dominan dibandingkan atribut lainnya sedangkan atribut daun bendera memiliki performa lebih rendah. Sementara, hasil akurasi algoritme C4.5 menggunakan pengujian *k-fold cross validation* akurasi tertinggi pada saat perhitungan *k-fold* 1 sebanyak 28,571%, sementara, *k-fold* 2 sebanyak 17,857%, *k-fold* 3 sebanyak 10,344%, *k-fold* 4 sebanyak 10,714%.

Saran untuk penelitian selanjutnya bagi perbaikan penelitian ini, dapat menggunakan algoritme yang berbeda untuk klasifikasi varietas unggul padi. Pada penelitian ini kelemahannya ada pada data yang memiliki ciri

nama yang sama dengan atribut yang lain dan apabila ingin membuat penelitian klasifikasi menggunakan algoritme C4.5 dengan memilih data yang cocok, agar hasil tree dibuat lebih optimal. Misalnya memilih data yang memiliki jumlah sama besar pada setiap kelasnya.

ex.php/publikasi/buku/item/668-deskripsi-varietas-unggul-baru-padi-2017> [Diakses 12 November 2018]

6. DAFTAR PUSTAKA

- Hermawati, F. A., 2013. *Data Mining*. Yogyakarta: CV ANDI OFFSET.
- Imanda, A.C., 2017. *Klasifikasi Kelompok Varietas Unggul Padi Menggunakan Modified K-Nearest Neighbor*. Skripsi Sarjana. Universitas Brawijaya.
- Indonesia, A.P.J.I., 2017. *Penetrasi & Perilaku Pengguna Internet Indonesia 2017*. [pdf] Asosiasi Penyelenggara Jasa Internet Indonesia. Tersedia di: <<https://apjii.or.id/survei2017/download/R0bygED5JOjKLZfTs8lw7dWF9qaCA1>> [Diakses 06 Januari 2019]
- Kamagi, D.H., dan Hansun, S., 2014. *Implementasi Data Mining Dengan Algoritma C4.5 Untuk Memprediksi Tingkat Kelulusan Mahasiswa*. [online] tersedia di: <<http://ejournals.umn.ac.id/index.php/TI/article/view/327/293>> [Diakses 07 Januari 2019]
- Kholifah, A.N., 2016. *Analisis Klasifikasi Pada Nasabah Kredit Koperasi X Menggunakan Decision Tree C4.5 Dan Naive Bayes*. S1. Universitas Negeri Yogyakarta. Tersedia di <<http://eprints.uny.ac.id/41356/>> [Tanggal Diakses : 07 Januari 2019]
- Nofriansyah, D., 2012. *Konsep Data Mining Vs Sistem Pendukung Keputusan*. [e-book] Tempat Publikasi : CV Budi Utama, Tersedia Melalui : <https://books.google.co.id/books?id=PoJyCAAQBAJ&pg=PA10&dq=data+mining+merupakan&hl=id&sa=X&ved=0ahUKEwinxZCUu_bVAhWLLo8KHbNnDV8Q6AEILTAB#v=onepage&q=data%20mining%20merupakan&f=false> [Tanggal Diakses : 26 Maret 2019]
- Wahab, M.I., Satoto., Rachmat, R., Guswara, A. & Suharna., 2017. *Deskripsi Varietas Unggul Padi*. [pdf] Badan Penelitian dan Pengembangan Pertanian. Tersedia di: <[Fakultas Ilmu Komputer, Universitas Brawijaya](http://bbpadi.litbang.pertanian.go.id/ind</p></div><div data-bbox=)