

PREDIKSI PANEN PADI MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOUR

Willmen TB Panjaitan^{1*}, Ema Utami², Hanif Al Fatta³

¹²³Magister Teknik Informatika, Fakultas Teknik Informatika, Universitas AMIKOM
Yogyakarta

Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta, 55283

*Email: willmen.panjaitan@students.amikom.ac.id

Abstrak

Krisis ketahanan pangan menjadi perhatian penting termasuk seluruh negara didunia. Berdasarkan laporan dari FAO, krisis ketahanan pangan ini telah berlangsung dan diperkirakan akan berlangsung lama apabila negara-negara tidak mempersiapkan strategi terhadap ketahanan pangan. Indonesia sebagai salah satu negara dengan luas pertanian yang memiliki potensial tentu telah mempersiapkan berbagai strategi untuk menyikapi masalah ketahanan pangan termasuk ketersediaan bahan makanan seperti beras. Pada penelitian ini akan dibahas mengenai penerapan salah satu metode datamining dalam proses prediksi panen padi. Dataset yang digunakan berasal dari BPS Provinsi D.I Yogyakarta dari tahun 2000 hingga 2017. Adapun metode yang digunakan yaitu k-Nearest Neighbor serta untuk pengujian nya digunakan RMSE. Hasil dari penelitian ini, metode k-Nearest Neighbor dengan model sederhana dapat melakukan prediksi terhadap panen padi dengan nilai RMSE 90532.279 +/- 0.000 dan Absolute Error 80876.833 +/- 40682.076.

Kata kunci: Absolute Error, Datamining, k-Nearest Neighbor, RMSE

1. PENDAHULUAN

Food and Agriculture Organization (FAO) melaporkan pada tahun 2016 situasi ketahanan pangan dunia memasuki keadaan yang menurun cukup drastis khususnya di beberapa wilayah negara sub-sahara Afrika, Asia Tenggara dan Asia Barat. Hal yang paling mempengaruhi keadaan tersebut adalah situasi konflik khususnya dimana terdapat konflik yang berdampak pada ketersediaan pangan serta dipengaruhi juga dengan kondisi kekeringan maupun banjir serta fenomena *El Nino* (FAO, 2017). Kegagalan dalam mengurai masalah ketahanan pangan dunia sangat berkaitan erat dengan peningkatan konflik dan kekerasan di beberapa bagian negara. Selama satu dekade terakhir, terjadi peningkatan konflik dan semakin kompleks sehingga sulit untuk dikendalikan (FAO, 2017). Salah satu yang menjadi perhatian penting adalah ketersediaan bahan makanan terutama bagi anak-anak serta dampak kekurangan gizi yang banyak ditemukan di negara-negara yang terkena dampak konflik. Hal ini cukup menjadi masalah yang perlu mendapatkan perhatian khusus karena hingga tahun 2030 diprediksi masalah ini akan terus terjadi (FAO, 2017).

Perencanaan ekonomi Indonesia mengikuti rencana pembangunan 20 tahun dari tahun 2005 hingga 2025, dibagi menjadi Rencana Jangka Menengah Lima Tahunan (RPJMN), masing-masing dengan prioritas pembangunan yang berbeda. RPJMN 2010–2014 memiliki fokus gizi dan termasuk target spesifik untuk mengurangi dari 37 menjadi 32 persen. Sementara RPJMN 2015-2019 menekankan pada pembangunan infrastruktur dan program bantuan sosial yang berkaitan dengan pendidikan dan kesehatan. RPJMN ini juga menyoroti dua tujuan yang berbeda untuk sektor pertanian yaitu untuk meningkatkan produksi beras untuk ketahanan pangan dan untuk mengembangkan tanaman bernilai lebih tinggi untuk meningkatkan mata pencaharian pedesaan (FAO, 2017). Badan Pusat Statistik (BPS) menjadi salah satu lembaga yang melakukan pencatatan terhadap seluruh data statistik yang berkaitan dengan beberapa hal seperti harga pangan, jumlah penduduk, jumlah sarana sekolah maupun data pangan. Ketersediaan data dapat dimanfaatkan lebih lanjut dengan menerapkan metode-metode data mining dalam pengolahannya.

Terdapat beberapa definisi berbeda yang diusulkan dalam berbagai literatur ilmiah sebagai pengertian dari data mining (Kuhkan, 2016). Beberapa definisi yang paling umum antara lain:

- a) Data mining meliputi deteksi pola valid, baru, dan mudah dipahami dalam set data; dengan kata lain, itu adalah proses yang ekstrak pengetahuan dari set data dengan menggunakan teknik cerdas.
- b) Data mining adalah bidang interdisipliner yang telah terintegrasi berbagai bidang seperti database, statistik, pembelajaran mesin dan bidang terkait lainnya, sehingga informasi berharga dan pengetahuan yang tersembunyi dalam jumlah besar data dapat diekstraksi

2. METODOLOGI

2.1 Literature Review

Pada bagian ini membahas mengenai study literature yang berkaitan dengan penelitian antara lain:

Djafar et al (Djafar dkk, 2017) melakukan penelitian mengenai produksi padi di wilayah Sulawesi Tenggara dengan menggunakan data dari tahun 1974 hingga 2014. Adapun metode yang digunakan pada penelitian tersebut adalah Fuzzy Timeseries dan menghasilkan peramalan produksi padi pada tahun 2015 dengan nilai 657768.25191 dengan nilai toleransi kesalahan MAPE 5.51%.

Asih et al (Asih dkk, 2017) melakukan penelitian mengenai prediksi produksi padi dengan menggunakan data Dinas Pertanian Tanaman Pangan dan Hortikultura Provinsi Kalimantan Barat, Dinas Pertanian dan Peternakan Kabupaten Kubu Raya dan data Badan Pusat Statistik Kalimantan Barat dari tahun 2009 hingga 2016 yang dibagi menjadi 3 periode yaitu Januari-April, Mei-Agustus dan September-Desember. Diperoleh hasil untuk batas bawah nilai pengujian sebesar 4.369.2(BPN-) dan batas atas nilai pengujian sebesar 19.744.3(BPN+).

Atik et al (Nurmasani, 2017) melakukan penelitian mengenai prediksi produksi padi menggunakan data dari BPS seluruh provinsi di Indonesia dari tahun 2006 hingga 2015. Algoritma yang digunakan yaitu SVM dengan kernel RBF dengan parameter yang telah ditentukan bersifat auto. Diperoleh hasil bahwa tahun 2007 merupakan prediksi yang memiliki nilai kesesuaian terbaik dengan nilai RMSE sebesar $1.20E+06$, R-square 0.794 dan Adjusted R-square 0.788 dimana nilai tersebut memiliki persentase kecocokan 78.8% [10]

Sellam dan Poovammal (Sellam and Poovammal, 2016) [15] melakukan penelitian mengenai prediksi panen padi dengan menggunakan analisis regresi. Adapun parameter yang disertakan antara lain luas area sawah, curah hujan serta indeks harga pangan. Adapun dataset yang digunakan pada penelitian ini bersumber dari departemen statistic dan agriculture darinational informatics center India dari tahun 2000 hingga 2011. Pada penelitian tersebut digunakan tools pendukung yaitu Matlab. Hasil penelitian tersebut menyebutkan bahwa nilai R2 sebesar 0.7 digunakan sebagai variable dalam perhitungan dimana Area under Cultivation (AUC), Annual Rainfall (AR) and Food Price Index (FPI) yang menghasilkan rata-rata 70% nilai pengaruh dari parameter tersebut dalam prediksi panen padi.

Jang dan Choe (Jang and Choe, 2014) melakukan penelitian mengenai prediksi produktivitas padi dengan menggunakan neural network. Penelitian ini merujuk pada penelitian sebelumnya yang telah dilakukan oleh the Ministry of Agriculture, Food and Rural Affairs (MAFRA) dari Korea Selatan dimana menitik beratkan pada factor temperature. Dataset mengenai factor temperature bersumber dari Korea Meteorological Administration (KMA) serta dataset berupa produktivitas panen diperoleh dari Korea National Statistical Office (KOSTAT). Dalam penelitian ini tidak digambarkan secara mendetail mengenai perhitungan algoritma yang digunakan namun pada tahap hasil dilakukan pengujian dengan menggunakan RMSE dan AIC untuk kombinasi factor yang menentukan dari prediksi produktivitas padi. Diperoleh hasil bahwa neuron ke-6 dengan nilai RMSE untuk training 0.07891 dan validation 0.08091 serta AIC -11114 merupakan model prediksi yang terbaik.

Dey et al (Dey dkk, 2017) melakukan penelitian mengenai prediksi produksi padi di Bangladesh. Dataset yang digunakan bersumber dari laporan tahunan agriculture Bangladesh. Terdapat parameter tambahan yang digunakan pada training set yaitu luas, curah hujan dan temperature. Sedangkan pada testing set digunakan untuk memprediksi dan proses validasi dilakukan dengan perhitungan nilai RMSE, MSE, MAE dan R-Square. Adapun algoritma yang digunakan untuk memprediksi yaitu SVM regression, Multiple Linear Regression dan Adaboost serta Nonlinear regression. Parameter yang digunakan pada penelitian tersebut antara lain curah hujan, temperature, kelembaban, luas wilayah serta hasil. Hasil yang didapat bervariasi untuk tiap daerah yaitu SVM Regression (SVMR) merupakan metode prediksi yang terbaik untuk memprediksi panen di daerah Aman Rice dengan RMSE 0.18, sedangkan MNR merupakan metode prediksi yang terbaik untuk memprediksi panen di daerah Aus dengan RMSE 0.427 dan Boro dengan RMSE 0.369.

2.2 Sumber Data

Sumber data pada penelitian kali ini yaitu menggunakan data yang diperoleh dari BPS Provinsi D.I Yogyakarta dalam publikasi berjudul Provinsi Yogyakarta Dalam Angka dalam kurun waktu tahun 2000-2017 dan dapat diakses melalui url: <https://yogyakarta.bps.go.id>. Data tersebut kemudian diolah sehingga menjadi dataset yang dapat digunakan dalam penelitian. Jenis data yang digunakan adalah sekunder karena bersumber dari pihak atau sumber lain dan kita hanya memerlukan beberapa langkah untuk menjadikan data tersebut siap untuk digunakan.

2.3 Metode Yang Diusulkan

Adapun metode yang dimaksud yaitu menggunakan k-nearest neighbor (k-NN atau KNN) dimana akan dibandingkan hasil produksi dengan jarak antar objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.

Algoritma KNN didasarkan pada pembelajaran dengan analogi yaitu dengan membandingkan contoh tes yang diberikan dengan contoh-contoh pelatihan yang mirip. Contoh pelatihan dijelaskan oleh atribut n . setiap contoh merupakan titik dalam ruang n -dimensi. Dengan cara ini, semua contoh pelatihan disimpan di ruang pola n dimensi. Ketika diberikan contoh yang tidak diketahui, algoritma KNN mencari ruang pola untuk contoh k pelatihan yang paling dekat dengan contoh yang tidak diketahui. Contoh k pelatihan ini adalah k "Nearest Neighbour" dari contoh diketahui "kedekatan" yang didefinisikan dalam hal jarak metrik, seperti pada jarak Euklidian.

Ide dari metode k-Nearest Neighbour 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 ke dalam kelas v . Jika kita mengasumsikan bahwa f adalah fungsi halus, sebuah ide yang masuk akal adalah untuk mencari sample dalam data training yang paling dekat (dalam hal variable independen) dan untuk menghitung v dari nilai-nilai y untuk sampel. Ketika kita membahas tentang Neighbour kita menyiratkan adanya jarak atau mengukur perbedaan yang dapat kita hitung antar sample berdasarkan pada variable independen. Measure of Distance yang paling populer adalah Euclidian Distance. Euclidian of Distance antar poin x dan u dapat dirumuskan sebagai (Leung, 2007):

$$D(x,u) = \sqrt{\sum_{i=1}^n (x_i - u_i)^2} \quad (1)$$

Root Mean Square Error (RMSE) telah digunakan sebagai metrik statistik standar untuk mengukur kinerja model meteorologi, kualitas udara, dan studi penelitian iklim. RMSE adalah metode alternatif untuk mengevaluasi teknik peramalan yang digunakan untuk mengukur tingkat akurasi hasil prakiraan suatu model (Chai and Draxler, 2014). Adapun persamaan dapat diperoleh dengan rumus:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (2)$$

3. HASIL DAN PEMBAHASAN

3.1 Metode Pengumpulan Data

Pengumpulan data primer pada penelitian kali ini yaitu bersumber dataset yang diperoleh dari BPS Provinsi Daerah Istimewa Yogyakarta dengan judul publikasi D.I Yogyakarta dalam angka. Adapun data yang digunakan sebanyak 51 data yaitu dari tahun 2000 hingga 2017. Data panen padi pertahunnya dibagi menjadi 3 bagian yaitu Quartal 1 (Januari hingga April), Quartal 2 (Mei hingga Agustus) dan Quartal III (September hingga Desember). Tabel 1 menunjukkan data mentah dari sumber yang kita gunakan, adapun contoh tersebut merupakan data hasil panen pada tahun 2015 (dalam ton)

Table 1. Dataset awal dari data BPS

Jenis Tanaman/ Crops	Sub Round / Sub Rounds			Jumlah/ Total
	I (Januari - April)	II (Mei - Agustus)	III (September - Desember)	
1	2	3	4	5
1. Padi Sawah / Wetland Paddy	313285	289687	143838	746810
2. Padi Ladang / Dryland Paddy	191965	6351	10	198326
3. Jagung/ Maize	204626	27168	67290	299084
4. Kedelai/ Soybeans	2789	14366	1667	18822
5. Kacang Tanah/ Peanuts	23096	49132	11072	83300
6. Kacang Hijau/ Mung Bean	92	111	27	230
7. Ubi Kayu/ Cassava	5354	755178	112830	873362
8. Ubi Jalar/ Sweet Potatos	1437	1182	3451	6070
9. Cantel/ Sorghum	37	7	7	51

3.2 Pengolahan Awal

Setelah mendapatkan kumpulan data, pengolahan awal dilakukan dimana data yang digunakan dalam penelitian ini hanya berupa field waktu (bulan panen) dan hasil panen. Untuk setiap tahunnya diperoleh 3 data untuk interval waktu panen ke-1 (Januari-April), panen ke-2 (Mei-Agustus) dan panen ke-2 (September-Desember). Pengolahan awal data yang dimaksud dengan pengolahan data dalam penelitian ini adalah proses pengelompokan data-data yang telah dikumpulkan sebelumnya dengan tujuan untuk menemukan variable-variabel yang akan digunakan beserta himpunan yang termasuk ke variable-variabel yang digunakan dengan merujuk jurnal yang ada. Adapun hasil dari preprocessing yang diharapkan adalah seperti pada table.2

Table 2. Dataset setelah preprocessing

Waktu	Hasil Panen
Jan-2000	366539
May-2000	214836
Sep-2000	81670
...	...
Sep-2016	141859

3.3 Model dan Metode Yang Diusulkan

Adapun metode yang dimaksud yaitu menggunakan k-nearest neighbor (k-NN atau KNN) dimana akan dibandingkan hasil produksi dengan jarak antar objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Algoritma KNN didasarkan pada pembelajaran dengan analogi yaitu dengan membandingkan contoh tes yang diberikan dengan contoh-contoh pelatihan yang mirip. Contoh pelatihan dijelaskan oleh atribut n . setiap contoh merupakan titik dalam ruang n -dimensi. Dengan cara ini, semua contoh pelatihan disimpan diruang pola n dimensi. Ketika diberikan contoh yang tidak diketahui, algoritma KNN mencari ruang pola untuk contoh k pelatihan yang paling dekat dengan contoh yang tidak diketahui. Contoh k pelatihan ini adalah k "Nearest Neighbour" dari contoh diketahui "kedekatan" yang didefinisikan dalam hal jarak metrik, seperti pada jarak Euklidean.

KNN merupakan algoritma non parametric lazy learning. Hal ini dikarenakan algoritma KNN tidak membuat asumsi apapun pada distribusi data pokok. Keuntungan ini karena mayoritas data praktis tidak mematuhi asumsi teoritis yang dibuat dan disinilah algoritma non parametric seperti KNN digunakan. KNN juga merupakan algoritma lazy learning dikarenakan tidak menggunakan generalisasi sehingga fase training sangat cepat. Kurangnya generalisasi artinya KNN menyimpan semua data training. KNN menghasilkan keputusan berdasarkan seluruh training dataset (Prajes dan Roy, 2014)

Algoritma kNN adalah salah satu algoritma klasifikasi yang paling terkenal digunakan untuk memprediksi kelas dari catatan atau (sampel) dengan kelas yang tidak ditentukan berdasarkan kelas dari catatan tetangganya. algoritma ini terbuat dari tiga langkah sebagai berikut (Gabralla dan Abraham, 2013):

- Menghitung jarak record masukan dari semua catatan pelatihan.
- Mengatur catatan pelatihan berdasarkan jarak dan pemilihan K-tetangga terdekat.
- Menggunakan kelas yang memiliki mayoritas diantara k-tetangga terdekat (metode ini menganggap kelas sebagai kelas record input yang diamati lebih dari semua kelas-kelas lain antar K-tetangga terdekat)

Classifier berasumsi jarak catatan dari satu sama lain sebagai kriteria untuk kedekatan mereka dan memilih catatan paling mirip. Ada banyak metode untuk menghitung jarak seperti fungsi jarak Euclidean, Manhattan, dll, di antaranya fungsi jarak Euclidean adalah salah satu yang paling umum didefinisikan sebagai Persamaan yaitu:

$$D(x_i, x_j) = \sqrt{\sum_{i=1}^n (a_r(x_i) - a_r(x_j))^2} \quad (3)$$

Keterangan:

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

(x_i) : Record ke-i

(x_j) : Record ke-j

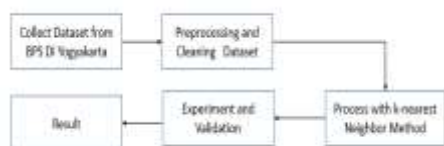
a_r = Data ke-r

$i, j = 1, 2, 3, \dots, n$

Secara umum, untuk memprediksi kelas rekor baru, algoritma mencari catatan serupa di antara set catatan pelatihan, sehingga jika catatan telah n atribut, maka akan menganggap mereka sebagai vektor dalam ruang n -dimensi dan memprediksi label kelas rekor baru berdasarkan kriteria jarak dalam ruang ini seperti jarak Euclidean serta label kelas dari tetangga

3.4 Eksperimen dan Pengujian

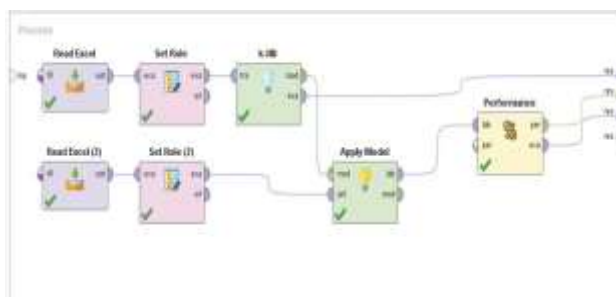
Adapun konsep dalam melakukan eksperimen dalam penelitian ini antara lain:



Gambar 1. Alur Eksperimen

Eksperimen dilakukan dengan menentukan nilai k dan penentuan metode untuk menentukan jarak(distance). Nilai k untuk menetapkan berapa k data dengan jarak terdekat yang akan dihitung atau diuji dalam percobaan ini. Nilai dari k biasanya adalah angka positif dan kecil. Dalam eksperimen ini, angka k yang diujikan berkisar dari 1 sampai dengan 10. Untuk measure_types akan dipilih numerical measures hal ini dikarenakan dataset dan data prediksi berupa angka.

Pada penelitian ini, digunakan tool rapidminer studio 8.0 dengan spesifikasi perangkat yaitu: Intel i5 2.5GHz, 8GB of RAM dan sistem operasi Windows 10. Adapun model yang diusulkan seperti pada gambar x dibawah:



Gambar 2. Model Yang Diusulkan

Table 3. Data perhitungan untuk 3 tahun terakhir dengan nilai k=1

No	Waktu	Hasil Panen	Prediksi Panen
1	Jan-2014	519178	357284
2	May-2014	274106	214149
3	Sep-2014	126289	80847
4	Jan-2015	505250	396111
5	May-2015	296038	203651
6	Sep-2015	143848	93236
7	Jan-2016	460500	379266
8	May-2016	280343	199490
9	Sep-2016	141859	91947

Table 4. Data perhitungan untuk 3 tahun terakhir dengan nilai k=3

No	Waktu	Hasil Panen	Prediksi Panen
1	Jan-2014	519178	255606
2	May-2014	274106	216607.25
3	Sep-2014	126289	192988.5
4	Jan-2015	505250	269180
5	May-2015	296038	224162.25
6	Sep-2015	143848	192347.25
7	Jan-2016	460500	262814.5
8	May-2016	280343	217548.25
9	Sep-2016	141859	199365.25

Table 5. Data perhitungan untuk 3 tahun terakhir dengan nilai $k=5$

No	Waktu	Hasil Panen	Prediksi Panen
1	Jan-2014	519178	200764
2	May-2014	274106	226450.875
3	Sep-2014	126289	240838.4167
4	Jan-2015	505250	209529
5	May-2015	296038	229545.5417
6	Sep-2015	143848	244016.875
7	Jan-2016	460500	205067.4167
8	May-2016	280343	232594.0417
9	Sep-2016	141859	250177.2083

Table 6. Data perhitungan untuk 3 tahun terakhir dengan nilai $k=7$

No	Waktu	Hasil Panen	Prediksi Panen
1	Jan-2014	519178	241154.7083
2	May-2014	274106	221488.7083
3	Sep-2014	126289	207984.1806
4	Jan-2015	505250	244244.5833
5	May-2015	296038	224312.6389
6	Sep-2015	143848	208917.7083
7	Jan-2016	460500	250261.8056
8	May-2016	280343	228270.4722
9	Sep-2016	141859	213167.7361

Table 7. Data perhitungan untuk 3 tahun terakhir dengan nilai $k=9$

No	Waktu	Hasil Panen	Prediksi Panen
1	Jan-2014	519178	223765.0687
2	May-2014	274106	222940.9188
3	Sep-2014	126289	223767.8313
4	Jan-2015	505250	225456.9625
5	May-2015	296038	224201.2875
6	Sep-2015	143848	229062.9687
7	Jan-2016	460500	231283.2125
8	May-2016	280343	229978.8625
9	Sep-2016	141859	225061.2813

Table 8. Hasil perhitungan RMSE dan Absolute error

Nilai K	Nilai RMSE	Nilai Absolute Error
1	90532.279 +/- 0.000	80876.833 +/- 40682.076
3	145655.237 +/- 0.000	115816.310 +/- 88329.103
5	179723.122 +/- 0.00	146230.145 +/- 104485.143
7	158515.731 +/- 0.000	127449.088 +/- 94254.798
9	169575.954 +/- 0.000	137147.331 +/- 99732.711

4. KESIMPULAN

Metode KNN dapat digunakan untuk melakukan prediksi panen padi. Adapun penerapan algoritma masih cukup sederhana dan untuk mendapatkan nilai RMSE dan absolute error yang baik, perlu dilakukan pengujian untuk nilai-K yang akan digunakan serta dilakukan perbandingan dengan metode lainnya sehingga didapat model terbaik. Penambahan faktor independen seperti curah hujan dan suhu dapat ditambahkan serta diterapkan komparasi dengan metode lainnya.

DAFTAR PUSTAKA

- A. Nurmasani, E. Utami, and H. Al Fatta, "Analisis Support Vector Machine," *J. Inf. Interaktif*, vol. 2, no. 1, 2017.
- Djafar, M. I. Sarita, and Y. P. Pasrun, "Peramalan jumlah produksi padi di sulawesi tenggara menggunakan metode," *semanTIK*, vol. 3, no. 2, pp. 113–120, 2017.
- FAO, IFAD, WFP, and WHO, "2017 Building Resilience for Peace and Food Security the State of Food Security and Nutrition in the World," *FAO, IFAD WFP*, pp. 1–54, 2017.
- FAO, "Socio-economic context and role of agriculture," 2017.
- I. H. Jang and Y. C. Choe, "Forecasting Rice Productivity Using a Neural Network Method: A Global Warming Scenario Impact Factors on Rice Productivity," *Adv. Sci. Technol. Lett.*, vol. 49, no. SoftTech, pp. 222–228, 2014.
- K. Asih, F. A. Setyaningsih, and D. M. Midyanti, "Aplikasi Prediksi Produksi Padi Menggunakan Regresi Interval Dengan Neural Fuzzy Di Kabupaten Kubu Raya," *Coding, Sistem Komput. Untan*, vol. 05, no. 2, pp. 108–118, 2017.
- K. Ming Leung, "k-Nearest Neighbor Algorithm for Classification," *NYU Polytechnic School of Engineering*, 2007.
- L. Gabralla and A. Abraham, "Computational Modeling of Crude Oil Price Forecasting: A Review of Two Decades of Research," *Mirlabs.Org*, vol. 5, pp. 729–740, 2013.
- M. Kuhkan, "A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm," *Int. J. Comput. Eng. Inf. Technol.*, vol. 8, no. 6, pp. 90–95, 2016.
- Prajesh P Anchalia, Kaushik Roy, "The k-Nearest Neighbor Algorithm Using MapReduce Paradigm", *Fifth International Conference on Intelligent System, Modeling and Simulation*, 2014.
- T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.
- U. K. Dey, A. H. Masud, and M. N. Uddin, "Rice Yield Prediction Model Using Data Mining," *Int. Conf. Electr. Comput. Commun. Eng. (ECCE)*, Febr. 16-18, 2017, pp. 321–326, 2017.
- V. Sellam and E. Poovammal, "Prediction of crop yield using regression analysis," *Indian J. Sci. Technol.*, vol. 9, no. 38, 2016.