Prediksi Cuaca di Kota Palembang Berbasis *Supervised Learning* Menggunakan Algoritma *K-Nearest Neighbour*

Alvi Syahrini Utami*[1], Dian Palupi Rini[2], Endang Lestari[3]
[1,2]Jurusan Teknik Informatika, Universitas Sriwijaya, Palembang
[3]Jurusan Sistem Informasi, Universitas Sriwijaya, Palembang
e-mail: *[1]alvisyahrini@ilkom.unsri.ac.id, [2]dprini@unsri.ac.id, [3]endang@ilkom.unsri.ac.id

Abstrak

Permasalahan cuaca yang dipengaruhi banyak faktor alam menyebabkan kondisi cuaca yang berubah - ubah sehingga kadang sulit diprediksi. Prediksi cuaca yang tepat diperlukan agar masyarakat dan para pengambil kebijakan dapat melakukan antisipasi terhadap hal ini. Banyaknya factor yang mempengaruhi cuaca menyebabkan kesulitan dalam mengklasifikasikan cuaca pada hari tertentu. Locality Sensitive Hashing (LSH) bekerja pada data pelatihan dengan memberikan nilai hash pada tiap vektor yang berisi nilai yang merepresentasikan faktor – faktor yang mempengaruhi cuaca dan melakukan pengklasifikasian cuaca. Untuk selanjutnya algoritma k-Nearest Neighbour (k-NN) yang akan menghitung prediksi terhadap faktor – faktor yang mempengaruhi cuaca pada suatu hari tertentu. Berdasarkan pengujian yang dilakukan, metode k-NN yang dihybrid dengan LSH dapat memprediksi nilai factor – factor yang mempengaruhi cuaca dengan cukup baik dengan nilai Mean Square Error (MSE) sebesar 0,301.

Kata kunci*—k-Nearest Neighbour (k-NN), prediksi cuaca, Locality Sensitive Hasihing (LSH)*

Abstract

Weather is influenced by many natural factors causing it to change frequently at any time so that it is sometimes difficult to predict. An accuratet weather prediction is needed so that people and policy makers can anticipate this problem. Many factors that influence the weather cause difficulty in classifying the weather on a particular day. Locality Sensitive Hashing (LSH) works on training data by assigning hash values to a vectors that contain values that represent factors that affect weather and perform weather classification. Furthermore, the k-Nearest Neighbor (k-NN) algorithm will calculate the predictions of the factors that affect the weather on a certain day. Based on the tests carried out, k-NN and LSH in weather prediction has Mean Square Error (MSE) 0,301.

Keywords— *k-Nearest Neighbou r(k-NN), weather forecasting, Locality Sensitive Hasihing (LSH)*

1. PENDAHULUAN

Prakiraan cuaca dapat diartikan sebagai suatu informasi kondisi udara yang akan terjadi di masa mendatang (paling lama 5 hari untuk daerah lintang sedang dan 2 hari untuk daerah tropis). Prakiraan cuaca memiliki peranan yang cukup penting, baik bagi para pelaku ekonomi maupun bagi para pengambil keputusan sehingga tidak akan menimbulkan kerugian moril maupun materil yang tinggi sebagai akibat dari kondisi cuaca di daerah tertentu.

Permasalahan cuaca yang dipengaruhi banyak faktor alam menyebabkan kondisi cuaca

yang berubah - ubah sehingga kadang sulit diprediksi. Kondisi alam kota Palembang yang memiliki banyak sungai dan rawa berpotensi menyebabkan banjir jika intensitas hujan cukup tinggi. Prediksi cuaca yang tepat diperlukan agar masyarakat dan para pengambil kebijakan dapat melakukan antisipasi terhadap hal ini. Aplikasi prakiraan cuaca diharapkan dapat menjadi salah satu alternatif solusi terhadap masalah ini.

Data yang digunakan untuk mendukung prakiraan cuaca biasanya berupa informasi yang dicatat setiap hari dalam jangka waktu tertentu, yang disebut dengan *time series*. *Time series* adalah serangkaian observasi $x_i(t)$; [$i = 1, \cdots, n$; $t = 1, \cdots, m$] yang dibuat secara sekuensial terhadap waktu, dimana indeks i adalah pengukuran yang dilakukan pada setiap waktu t [1]. Jika $n = 1$ disebut *univariate time series*, dan jika $n > 1$ disebut *multivariate time series*.

Suatu *time series* dapat digunakan sebagai data untuk diproses oleh suatu algoritma klasifikasi. Salah satu algoritma yang dapat digunakan untuk mengklasifikasikan data adalah *k-Nearest Neighbour* (*k-NN*). *k-NN* mengklasifikasikan data berdasarkan sejumlah k tetangga terdekat (*nearest neighbour*).

Beberapa penelitian yang terkait dengan prediksi cuaca dan metode *k-NN* telah cukup banyak dilakukan. Penelitian - penelitian tersebut antara lain dilakukan oleh Barded an Patole dalam [2] yang membandingkan antara Jaringan syaraf Tiruan, *k-NN*, dan Algoritma *Naive Bayes* untuk mengklasifikasikan dan memprediksi cuaca dengan hasil akurasi terbaik didapatkan oleh metode *k-NN*. Penelitian lainnya mengkaji metode *k-NN* untuk *multivariate time series* dilakukan oleh [3]. Hasil penelitian menunjukkan bahwa kinerja *k-NN* sangat baik untuk mengklasifikasi *dataset time series*.

Pada penelitian ini peneliti juga mempergunakan metode *Locality Sensitive Hashing* (*LSH*) sebagai salah satu metode yang di harapkan mampu memberikan solusi lebih baik untuk mempersingkat waktu perhitungan pada *k-NN* dan juga mempergunakan teori optimasi dalam menentukan solusi terbaik yang di butuhkan oleh sistem. Algoritma *k-NN* yang di*hybrid* dengan *LSH* diharapkan akan memberikan akurasi yang baik untuk memprediksi cuaca di kota Palembang.

Penelitian [4] membandingkan tiga buah metode untuk prakiraan cuaca. Metode - metode tersebut antara lain *Autoregressive Integrated Moving Average* (ARIMA), *Neural Network* atau jaringan syaraf, dan *Adaptive Splines Threshold Regression* (ASTAR). Ketiga metode tersebut menghasilkan nilai prakiraan tiga unsur cuaca, yaitu suhu, kelembeban nisbi, dan curah hujan. Ketiga metode dievaluasi dengan nilai korelasi dan *Root Mean Square Error* (RMSE). Hasil penelitian menunjukkan bahwa metode ASTAR menghasilkan prakiraan yang lebih baik. Penelitian tentang prakiraan cuaca juga dilakukan oleh [5] dengan menggunakan metode untuk *data mining* yaitu *K-medoids* dan algoritma *Naive Nayes*. Prakiraan cuaca dilakukan berdasarkan parameter - parameter suhu, kelembaban, dan angin. Kesimpulan pada penelitian ini menunjukkan bahwa prakiraan cuaca dengan menggunakan teknik *data mining* memberikan hasil prediksi yang cukup baik.

Berikutnya adalah penelitian - penelitian yang terkait dengan metode yang akan digunakan yaitu k-NN. Salah satu penelitan tersebut adalah [6] mengimplementasikan metode *k-NN* untuk menentukan lokasi pos hujan terdekat dengan rute perjalanan untuk aplikasi Clearroute. Clearroute sendiri adalah apli- kasi yang dibangun untuk memudahkan seseorang mengetahui keadaan cuaca pada rute yang akan dilaluinya. Penghitungan akurasi dengan menggunakan *confussion matrix* dan nilai $k = 3$ menghasilkan akurasi sebesar 73%.

Locality Sensitive Hashing (*LSH*) adalah metode yang telah dikenal di dunia komputasional untuk pencarian tetangga terdekat yang efisien pada *dataset* berdi- mensi tinggi. [7] menggunakan *LSH* untuk mendeteksi gempa bumi dengan studi kasus pengskalaan *data-driven science*. Sedangkan [8] menggunakan *LSH* dalam *data-driven predictive conrtol* yang menghasilkan simulasi berupa grafik, perhitungan error, dan waktu komputasi.

Berdasarkan hasil penelitan terdahulu, metode *k-NN* dan *LSH* memiliki tingkat akurasi yang cukup baik. Dalam penelitian ini *Hybrid* metode *k-NN* dan *LSH* akan digunakan pada untuk memprediksi cuaca di kota Palembang. Setelah implementasi dilakukan, akan dianalisis hasil prediksi terhadap data sebenarnya yang akan menjadi tingkat akurasi dari metode ini. Tingkat akurasi akan dihitung menggunakan Mean Square Error.

2. METODE PENELITIAN

Pada penelitian ini digunakan metode *LSH* untuk menghasilkan nilai *hash* untuk masing – masing *record* pada data pelatihan. Nilai *hash* yang didapatkan kemudian akan digunakan untuk mengklasifikasikan data uji yang akan dihitung menggunakan *k-NN*. Data yang digunakan pada penelitian ini diperoleh dari Badan Meteorologi Klimatologi dan Geofisika (BMKG) kota Palembang. Data berupa fitur - fotur yang mempengaruhi cuaca seperti suhu minimum, suhu maksimum, suhu rata - rata, kelembaban udara, lama penyinaran, arah angin, kecepatan angin, kecepatan angin terbesar, dan curah hujan.

2.1 Tahapan penelitian

Adapun tahapan penelitian yang dilakukan adalah sebagai berikut :

1. Pengumpulan data yang berupa data yang mempengaruhi cuaca kemudian data yang tidak sesuai dengan format yang diinginkan akan dibuang. Data yang telah melalui proses penyaringan data akan dipilih untuk masuk ke *dataset* dan akan digunakan sebagai data latih dan data uji pada penelitan.

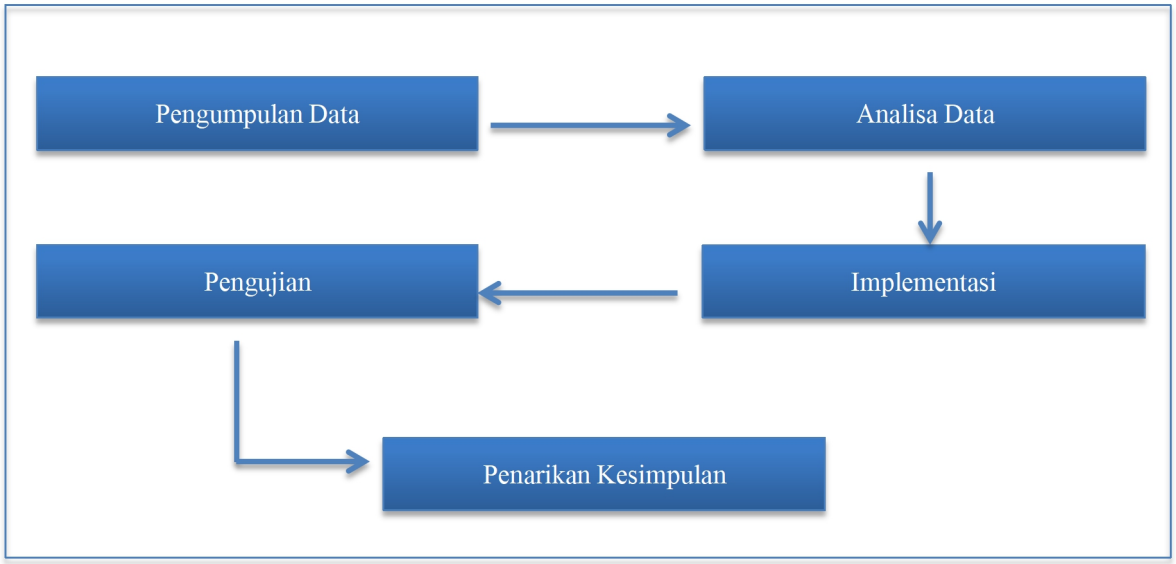
2. Analisa data dilakukan untuk mengidentifikasi dan mengevaluasi permasalahan atau anomali yang mungkin saja ada pada data mentah yang dikumpulkan. Untuk kasus seperti ini dibutuhkan suatu praproses yang akan menghasilkan data yang sudah normal untuk diproses menggunakan metode yang dipilih. Pada tahap ini juga dilakukan pemilihan metode yang sesuai dengan data yang diperoleh maupun hasil analisa data.

3. Implementasi menggunakan metode k-NN yang di*hybrid* dengan *LSH*. Pada tahap ini terlebuh dulu dilakukan perancangan yang akan dibuat berupa *flowchart*, diagram alir data, dan lain - lain. Kemudian akan dilanjutkan dengan pembuatan kode program sesuai dengan rancangan yang ada.

4. Kode program yang telah selesai dibuat akan diuji menggunakan *dataset* yang telah ada. *Dataset* akan dibagi menjadi data pelatihan dan data pengujian. Pada tahap ini juga akan dilakukan analisa terhadap hasil dari metode yang dipilih. Pengukuran terhadap perfomansi metode dilihat dari akurasi yang dihasilkan. Penghitungan akurasi menggunakan Mean Square Error (MSE).

5. Penarikan kesimpulan berdasarkan analisa hasil dan perhitungan akurasi.

Tahapan peneltian ini dapat dilihat pada gambar 2.1.



Gambar 2.1. Tahapan Penelitian

2. 2 Langkah – langkah implememtasi

Garis besar tahap-tahap implementasi metode *k-NN* dan *LSH* pada penelitian ini adalah sebagai berikut :

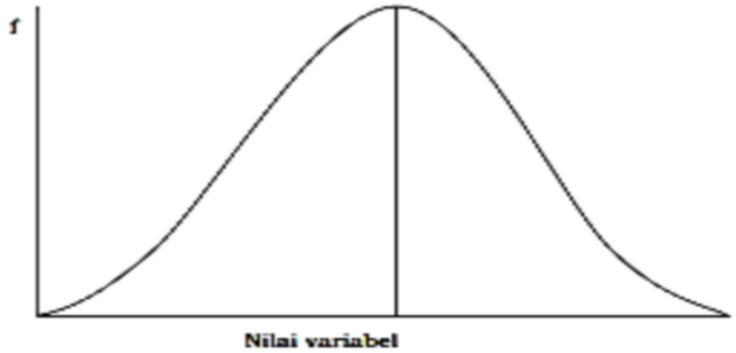
Inisialisasi. Menentukan nilai-nilai variabel *n*,*m*,*A* and *a*. Dalam hal ini, *m* dan *n* adalah ukuran baris dan kolom matriks yang akan dibuat berdasarkan jumlah fitur dan *record* data. *A* adalah matriks yang terbentuk, dan *a* adalah vektor yang ukurannya tergantung dengan jumlah fitur pada dataset. Proses penentuan nilai-nilai variabel berdasarkan data yang didapatkan dari Badan Meteorologi, Klimatologi dan Geofisika (BMKG) Palembang. Jumlah data yang digunakan adalah 1200 *record* data yang akan dibagi menjadi data pelatihan dan data pengujian. Pada tahap ini juga dilakukan praproses data.

Langkah 1. Sebuah vektor *a* akan dibentuk dari banyaknya jumlah fitur, pada penelitian ini vector *a* berukuran 9. Pembentukan matriks *A* yang berukuran *m* x *n* dimana nilai *m* dan *n* berdasarkan nilai pada inisialisasi. Berdasarkan data yang didapatkan, terdapat 1200 record data dan masing – masing data memiliki 9 fitur sehingga matriks yang terbentuk berukuran 9 x 1200. Berikut ini desain vector a dan matriks A.

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

Langkah 2. Normalisasi data dengan tujuan data dapat terdistribusi normal. COntoh data berdistribusi normal mengikuti kurva pada gambar 2.1.



Gambar 2.1. Kurva Normal

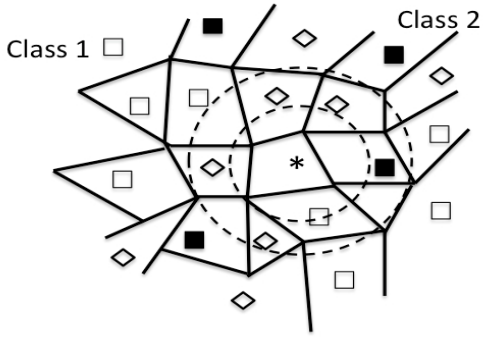
Langkah 3. Klasifikasi data dalam matriks menggunakan metode *LSH*, dimana pada langkah ini digunakan formula sebagai berikut [9] :

$$h : \Re^d \rightarrow Z, h_{a,b}(v) = \left\lfloor \frac{a^T v + b}{r} \right\rfloor$$

Dimana r > 0 adalah sebuah parameter dari fungsi *hash, a* $\in R^d$ dipilih secara bebas dari sebuah distribusi normal, *b* bernilai skalar dari sebuah ditribusi seragam (*uniform distribution*) dan $\lfloor x \rfloor$ adalah sebuah fungsi *floor,* dan *v* adalah vector yang berisi nilai fitur – fitur pada suatu *record* data.

Langkah 4. Menentukan data pengujian dan mencari prediksi fitur – fitur yang ada pada data pengujian dengan cara membandingkannya dengan data pelatihan. Prediksi cuaca pada *t + 1* dilakukan dengan menggunakan data saat *t* dan dicari jarak terdekatnya dengan data latih yang memiliki nilai *hash* yang sama atau bernilai +/- 1 terhadap nilai *hash* pada *t*. Misal didapat *s* adalah data terdekat dengan *t*, maka diambil *s + 1* sebagai prediksi untuk *t + 1*. Lakukan kembali normalisasi dan klasifikasi pada data uji seperti pada langkah 2 dan 3.

Langkah 5. Setelah didapatkan hasil klasifikasi terhadap vdata uji, dilakukan penghitungan jarak terpendek dengan metode *k-NN*, khususnya yang memiliki hasil klasifikasi yang sama antara data latih dan data uji. Kelas yang diambil adalah kelas yang terbanyak dan terdekat dengan data uji. Ilustrasi metode ini dapat dilihat pada gambar 2.2.

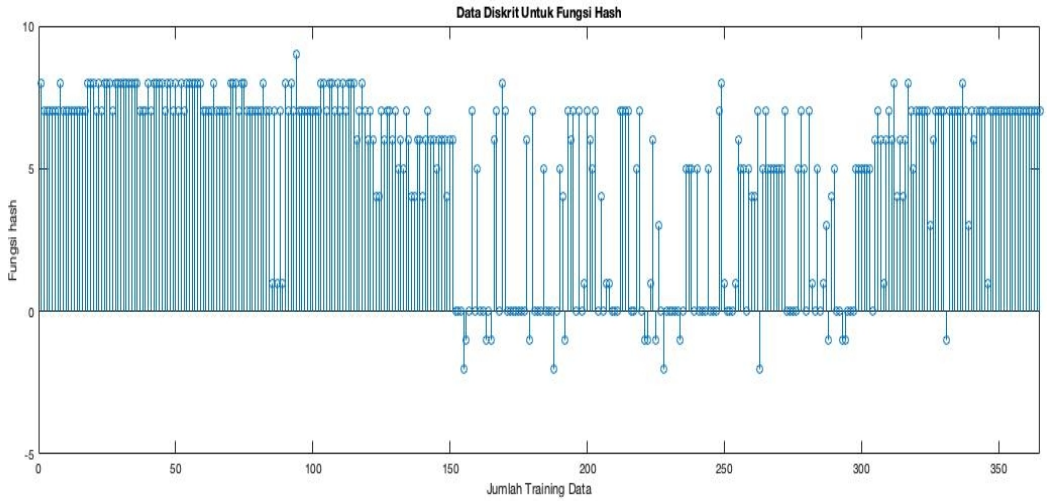


Gambar 2.2. Klasifikasi *k-Nearest Neighbour* [8]

3. HASIL DAN PEMBAHASAN

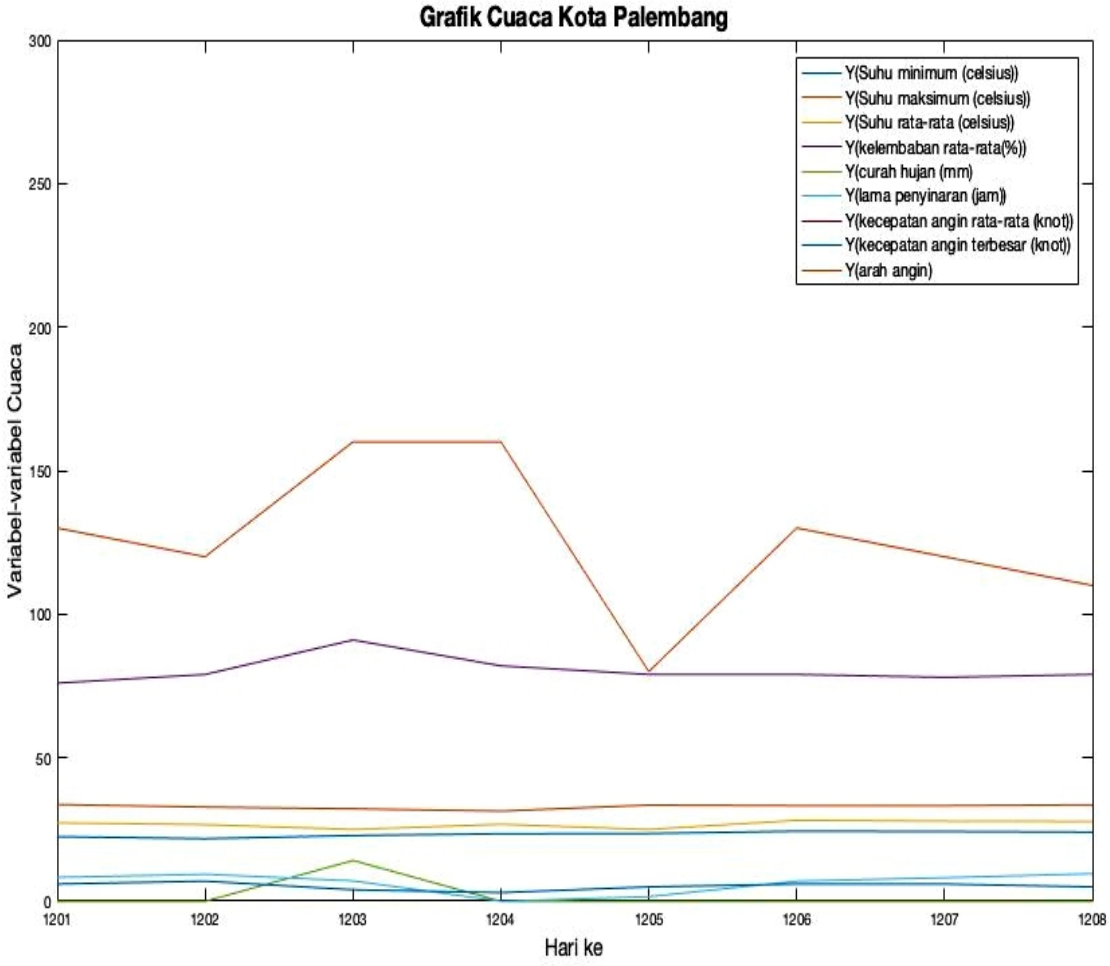
Sebelum digunakan, data akan dianalisis terlebih dahulu. Analisis ini dilakukan untuk memastikan bahwa data yang akan diproses oleh kedua metode ini benar – benar telah bebas dari *noise* atau anomali lainnya. *Noise* atau anomali yang ditemukan pada data mentah antara lain adalah adanya nilai yang kosong dan nilai yang besarnya anomali terhadap data lainnya. Untuk fitur arah angin, data mentah yang diperoleh bukan berupa angka, seperti barat, barat daya, selatan, dst., sehingga untuk fitur ini diberikan nilai angka sesuai dengan besarnya sudut arah angin yang titik nolnya dihitung dari utara dan searah jarum jam. Akibat dari hal ini, nilai – nilai pada fitur arah angin ada pada kisaran 0 – 360 dimana perubahan arah angin yang sedikit akan menyebabkan perubahan nilai yang cukup besar yang akan berpengaruh terhadap hasil prediksi.

Metode *LSH* memberikan nilai *hash* pada masing – masing *record* data yang digunakan untuk mengklasifikasikan data tersebut. Grafik nilai *hash* untuk sampel 300 data dari 1200 data dapat dilihat pada gambar 3.1. Nilai – nilai *hash* inilah yang akan menentukan hasil klasifikasi terhadap data uji.



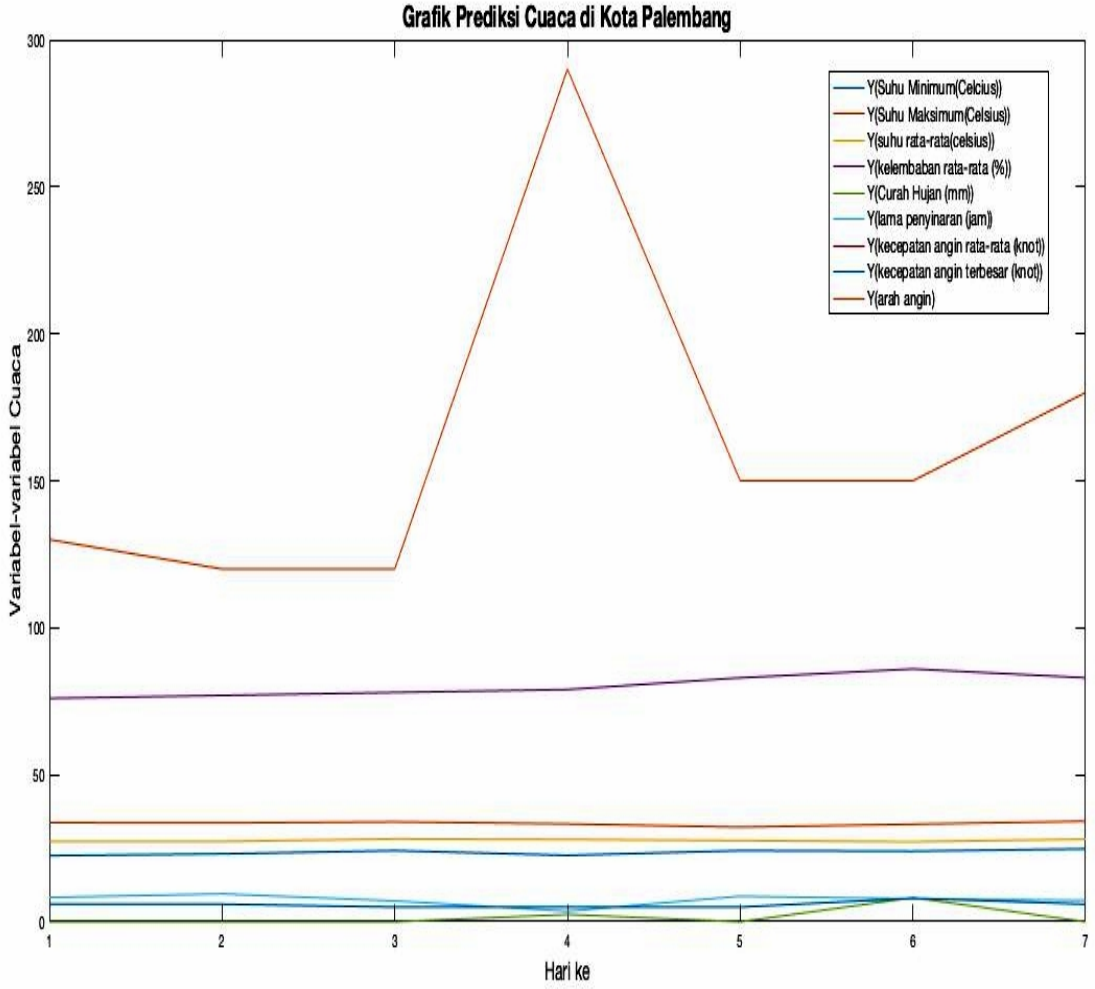
Gambar 3.1 Nilai *hash* 300 data

Setelah didapatkan nilai *hash* untuk masing – masing *record* data, pada data untuk pengujian yang dipilih akan dihitung prediksi nilai dari masing – masing fitur cuaca. Perhitungan ini akan berdasarkan nilai fitur yang terjadi pada hari sebelumnya dan nilai hash yang di dapat pada saat pelatihan. Pengujian dilakukan terhadap data ke-1201 sampai data ke-1207. Prediksi dilakukan terhadap nilai fitur – fitur pada tiap data uji tersebut. Gmabar 3.2 menunjukkan nilai – nilai fitur yang sebenarnya pada data uji, sedangkan hasil prediksi dapat dilihat pada gambar 3.3.



Gambar 3.2. Fitur Cuaca

Berdasarkan tampilan dari kedua grafik tersebut, hasil prediksi untuk fitur suhu minimum, suhu maksimum, suhu rata - rata, kelembaban udara, lama penyinaran, kecepatan angin, kecepatan angin terbesar, dan curah hujan mendekati nilai fitur yang sebenarnya. Sedangkan untuk fitur arah angina memang ada perbedaan yang cukup signifikan pada grafik, hal ini disebabkan karena nilai pada fitur arah angin menggunakan satuan derajat yang mengakibatkan jika ada sedikit saja perbedaan arah angin akan menyebabkan perubahan nilai yang cukup besar. Untuk mengatasi hal ini hasil uji juga akan dinormalisasi sehingga nilai – nilai yang didapatkan akan terdistribusi secara normal.



Gambar 3.2. Prediksi Fitur Cuaca

Pada tabel 3.1 menunjukkan nilai – nilai fitur hasil normalisasi. Nilai – nilai pada tabel ini akan digunakan untuk mengukur akurasi menggunakan Mean Square Error, dimana

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

dimana :

MSE : Mean Square Error
y : data sebenarnya
$\hat{y}$: data hasil prediksi
n : jumlah data uji

Tabel 3.1. Normalisasi Fitur Data Uji

Fitur	Hari ke-													
	1		2		3		4		5		6		7	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R
F1	-0.279	-0.279	-0.263	-0.298	-0.230	-0.327	-0.319	-0.247	-0.266	-0.142	-0.312	-0.243	-0.281	-0.235
F2	-0.016	-0.016	-0.0002	-0.023	0.012	-0.149	-0.203	-0.098	-0.103	0.172	-0.122	-0.036	-0.118	-0.011
F3	-0.167	-0.167	-0.158	-0.175	-0.131	-0.287	-0.260	-0.185	-0.197	-0.097	-0.246	-0.155	-0.223	-0.145
F4	0.975	0.975	1.087	1.129	1.103	0.973	0.290	0.855	0.928	1.620	0.967	1.025	0.724	1.104
F5	-0.807	-0.807	-0.842	-0.842	-0.830	-0.493	-0.536	-0.690	-0.757	-0.893	-0.639	-0.811	-0.709	-0.841
F6	-0.612	-0.612	-0.604	-0.607	-0.654	-0.629	-0.524	-0.690	-0.581	-0.842	-0.647	-0.647	-0.583	-0.637
F7	-0.666	-0.666	-0.692	-0.667	-0.706	-0.688	-0.509	-0.633	-0.656	-0.734	-0.643	-0.671	-0.605	-0.692
F8	-0.666	-0.666	-0.692	-0.667	-0.706	-0.688	-0.509	-0.633	-0.656	-0.734	-0.643	-0.671	-0.605	-0.692
F9	2.241	2.241	2.165	2.152	2.144	2.291	2.573	2.325	2.289	1.652	2.288	2.211	2.399	2.151

Keterangan :

F1 : suhu minimum (celsius)
F2 : suhu maksimum (celcius)
F3 : suhu rata – rata (celcius)
F4 : kelembaban rata – rata (%)
F5 : curah hujan (mm)
F6 : lama penyinaran (jam)
F7 : kecepatan angin rata – rata (knot)
F8 : kecepatan angin terbesar (knot)
F9 : arah angin (derajat)
P : nilai prediksi
R : nilai sebenarnya

Tingkat kesalahan pada prediksi cuaca menggunakan *LSH* dan *k-NN* didapat dari perhitungan *MSE* yang menghasilkan nilai sebesar 0,301.

4. KESIMPULAN

Prediksi cuaca di kota Palembang dengan menggunakan hybrid metode LSH dan k-NN telah dilakukan dan dapat berjalan dengan baik. Nilai *MSE* yang diperoleh sebesar 0,301 sehingga tingkat keakuratan dari hybrid metode ini adalah sekitar 70%.

5. SARAN

Untuk mendapatkan hasil akurasi yang lebih baik dapat digunakan jumlah data pelatihan dan pengujian yang lebih banyak atau menggunakan metode klasifikasi lain.

UCAPAN TERIMA KASIH

Penelitian ini dibiayai oleh Anggaran DIPA Badan Layanan Umum Universitas Sriwijaya Tahun Anggaran 2020 No. SP DIPA-023.17.2.677515/2020 Revisi ke 01 tanggal 16 Maret 2020 Sesuai dengan SK Rektor Nomor : 0684/UN9/SK.BUK.KP/2020 Tanggal 15 Juli 2020 dan kontrak nomor : 0163.268/UN9/SB3.LPPM.PT/2020.

DAFTAR PUSTAKA

[1] Yang, S. dan Shahabi, C. (2007). An efficient k-nearest neighbour search for multivariate time series. *Information and Computation*, 205:65–98.

[2] Barde, N. dan Patole, M. (2014). Classification and forecasting of waether using ann, k-nn, and naive bayes algorithms. *International Journal of Science and Research (IJSR)*, pages 1740 – 1742.

[3] Yang, S. dan Shahabi, C. (2007). An efficient k-nearest neighbour search for multivariate time series. *Information and Computation*, 205:65–98.

[4] Sutikno, Bekti, R. D., Susanti, P., dan Istriana (2010). Perkiraan cuaca de- ngan metode autoregressive integrated moving average, neural network, dan adaptive splines threshold autoregression di stasiun juanda surabaya. *Jurnal Sains Dirgantara*, 8(1):43 – 61.

[5] Biradar, P., Ansari, S., Paradkar, Y., dan Lohiya, S. (2017). Weather predic- tion using data mining. *International Journal of Engineering Development and Research (IJEDR)*, 5(2):211 – 214.

[6] Rosidi, A., Ginardi, R. H., dan Munif, A. (2017). Implementasi metode k- nearest neighbour untuk penentuan lokasi pos hujan terdekat dengantitik rute perjalanan pada aplikasi clearroute. *Jurnal Teknik ITS*, 6(2):A392 – A395.

[7] Rong, K., Yoon, C. E., Bergen, K. J., dan Elezabi, H. (2018). Locality sensitive hashing for earthquake detection : A case study of scalling data-driven science. In *The 44th International Conference on Very Large Data Bases*, volume 44, pages 1674 – 1687, Rio de Janeiro, Brazil.

[8] Saputra, H., Malyan, A. B. J., Supani, A., dan Indarto (2019). Data-driven predictive control menggunakan algoritma nearest neighbour untuk sistem yang tidak stabil. *Jurnal JUPITER*, 10(1):41 – 51.

[9] Datar, M., Immorlica, N., Indyk, P., dan Mirrokni, V. (2004). Locality- sensitive hashing scheme based on p-stable distribution. Number 20, pages 253–262, Brooklyn, New York, USA. ACM.