Jurnal Riset Rumpun Ilmu Teknik (JURRITEK)
Vol.1, No.1 April 2022
e-ISSN: 2829-016X; p-ISSN: 2829-0178, Hal 88-97

IMPLEMENTASI METODE K-NEAREST NEIGHBOUR DALAM MEMPREDIKSI CURAH HUJAN DI KOTA BOGOR

Naufal Rasyid[a],
[a] Program Studi Ilmu Komputer, , STIMIK ESQ
naufal.r@students.esqbs.ac.id

Trevy Jonatya Novella[b] ,
b Program Studi Ilmu Komputer, STIMIK ESQ
t.jonatya.n@students.esqbs.ac.id,

Ahlijati Nuraminah[c]
[c] Program Studi Ilmu Komputer, STIMIK ESQ
ahlijati.nuraminah@esqbs.ac.id

ABSTRAK

Accurate weather prediction information is important for various fields that are closely related to weather forecasting, such as agriculture, fisheries and many more. Because precise weather forecasts are very useful for various fields of carrying out various activities. Because of that, it is necessary to make an application to find weather or rainfall prediction information, so that the information can be utilized optimally by the community. In this journal the authors apply the k-nearest neighbors (k-NN) method based on rainfall data obtained from the Bogor climatology station from 2016-2017 and the test results show that the predicted rainfall for the Bogor area with the K-Nearest Neighbor algorithm obtained a value of 0, 93148.
Keywords*: K-Nearest Neighbor, Rainfall¸ Bogor City, Machine Learning*.

Abstrak

Informasi prediksi cuaca yang akurat penting untuk berbagai bidang yang erat kaitannya dengan prakiraan cuaca, seperti pertania, perikanan dan masih banyak lagi. Karena prakiraan cuaca yang tepat sangat bermanfaat untuk berbagai bidang melaksanakan berbagai kegiatan. Karena dari itu perlu dibuat suatu aplikasi untuk mengetahui informasi prediksi cuaca atau curah hujan, sehingga informasi dapat dimanfaatkan secara maksimal oleh masyarakat. Pada jurnal ini penulis menerapkan metode k-nearest neighbors(k-NN) berdasarkan data curah hujan yang didapakan dari stasiun klimatologi bogor dari tahun 2016-2017 dan hasil pengujian menunjukkan bahwa prediksi curah hujan daerah bogor dengan algoritma K-Nearest Neighbor didapatkan Nilai 0,93148.
Kata Kunci: K-Nearest Neighbour, Curah Hujan¸ Kota Bogor, Machine Learning

1. PENDAHULUAN

Dalam menghadapi permasalahan yang ada manusia banyak dibantu oleh teknologi, karena banyaknya metode yang dapat digunakan untuk menganalisis, mengklasifikasikan, dan memvisualisasikan objek yang dapat bermanfaat bagi kehidupan manusia. Prediksi adalah proses memperkirakan secara sistematis yang mungkin terjadi di masa yang akan datang berdasarkan informasi yang dimiliki dari masa lalu dan sekarang, agar kesalahannya dapat

Received Febuari 07, 2022; Revised Maret 2, 2022; April 22, 2022
*Corresponding author, e-mailnaufal.r@students.esqbs.ac.id

diperkecil. Prediksi tidak harus memberikan jawaban secara pasti kejadian yang akan terjadi, tetapi berusaha mencari jawaban sedekat mungkin dengan suatu hal yang akan terjadi.

Dikarenakan negara Indonesia yang dilintasi garis khatulistiwa, dua samudra dan benua, sehingga menjadikan Indonesia memiliki berbagai iklim yang dipengaruhi oleh matahari, curah hujan, suhu, kelembapan, suhu udara, tekanan udara dan angin. Dari setiap daerah dapat mengalami perbedaan cuaca maupun iklim, karena perbedaan ketinggian, daerah tekanan, arus laut, lintang, permukaan tanah. Salah satu yang penting untuk diketahui yaitu curah hujan karena dapat berdampak besar untuk kehidupan manusia.

Dengan memprediksi curah hujan kita dapat mengetahui dan memahami pengelolaan saluran pembuangan, pengelolaan air, pencegahan banjir, dan masih banyak lagi.Maka dari itu dibutuhkan metode yang dapat memprediksi curah hujan berdasarkan pola yang terjadi setelah perubahan iklim. Dengan adanya metode ini seluruh bencana yang berkaitan dengan cuaca dan perubahan iklim dapat diminimalisir karena adanya informasi akurat yang diperoleh. Pada penelitian ini penulis akan membahas model klasifikasi K-Nearest Neighbour yang akan digunakan untuk memprediksi curah hujan dan fokus terhadap curah hujan daerah Kota Bogor.

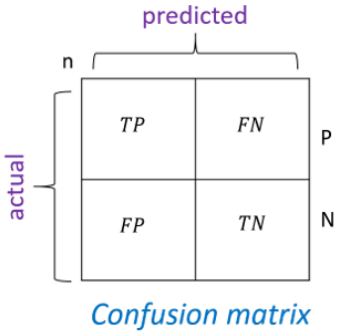
1. TINJAUAN PUSTAKA

1.1. K-Nearest Neighbor

K-Nearest Neighbours adalah salah satu algoritme klasifikasi paling dasar namun penting dalam Machine Learning. K-Nearest Neighbours adalah supervised learning domain dan intens dalam pengenalan pola, penggalian data, dan deteksi intrusi.Ini digunakan secara luas dalam skenario kehidupan nyata karena non-parametrik, yang artinya tidak membuat asumsi mendasar apa pun tentang distribusi data (berbeda dengan algoritme lain seperti Gaussian Mixture Models, yang mengasumsikan distribusi Gaussian dari data yang diberikan). Tujuan dari algoritma K-NN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan training samples. Teknik ini sangat sederhana dan mudah diimplementasikan.

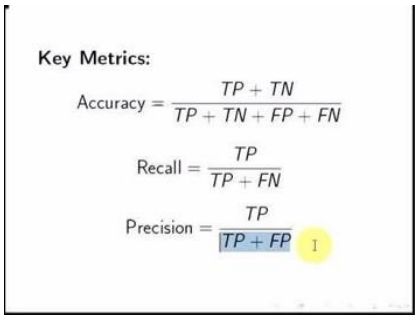
2.2 Confusion Matrix

Confusion Matrix adalah pengukuran performa untuk masalah klasifikasi machine learning dimana output dapat berupa dua kelas atau lebih. Confusion Matrix adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan nilai aktual. Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix yaitu True Positif, True Negatif, False Positif, dan False Negatif.



Gambar 1. Confusion matrix

Untuk menghitung confusion matrix bisa menggunakan rumus dibawah ini



Gambar 2. Rumus confusion matrix

Accuracy adalah kedekatan pengukuran dengan nilai tertentu. akurasi yang rendah menyebabkan perbedaan antara hasil dan nilai "sebenarnya". Akurasi yang tinggi membutuhkan presisi tinggi dan kebenaran yang tinggi. Presisi adalah sekumpulan titik data dari pengukuran berulang dengan besaran yang sama, himpunan tersebut dapat dikatakan akurat jika rata-ratanya mendekati nilai sebenarnya dari besaran yang diukur, sedangkan himpunan tersebut dapat dikatakan tepat jika nilainya dekat satu sama lain. Recall adalah bagian dari dokumen relevan yang diambil dibandingkan dengan jumlah total dokumen yang relevan (TP dibagi dengan TP+FN).

2.3 Principal Component Analysis (PCA)

PCA pada dasarnya adalah teknik reduksi dimensi sederhana yang mengubah kolom dari kumpulan data menjadi fitur set baru yang disebut Principal Components(PC). Informasi yang ada dalam kolom adalah jumlah varians yang didalamnya. Tujuan utama dari Principal Components adalah untuk merepresentasikan informasi dalam dataset dengan kolom seminimal mungkin

2. METODOLOGI PENELITIAN

3.1 Data

Pada penelitian ini data yang digunakan adalah data yang dikumpulkan dari https://www.bmkg.go.id/ berisi data curah hujan, yang terdiri dari variabel di bawah ini:

Tabel 2. Variabel data curah hujan

Variabel	**Keterangan**
Stasiun	Stasiun Cuaca
Tanggal	Tanggal
Tn	Suhu minimum (°C)
Tx	Suhu maksimum (°C)
Tavg	Suhu rata-rata (°C)
RR	Curah hujan (mm)
Hari_Hujan	Kebenaran dasar hujan hari ini
Besok_Hujan	Kebenaran dasar ramalan cuaca besok

3.2 Method

K-Nearest Neighbor (KNN) adalah metode non-parametrik yang digunakan untuk klasifikasi dan regresi. Input terdiri dari k training terdekat di ruang fitur. Outputnya tergantung pada apakah k-NN digunakan untuk klasifikasi atau regresi.Dalam klasifikasi k-NN, outputnya adalah keanggotaan kelas. Sebuah objek diklasifikasikan dengan pluralitas dari neighbors, dengan objek yang ditugaskan ke kelas yang paling umum di antara k neighbors terdekatnya (k adalah bilangan bulat positif, biasanya kecil). Jika k = 1, maka objek tersebut ditetapkan ke kelas neighbors tunggal terdekat.

Dalam regresi k-NN, outputnya adalah nilai properti objek. Nilai ini merupakan rata-rata nilai k neighbors terdekat. k-NN adalah jenis instance-based learning, atau lazy learning, di mana fungsinya hanya diperkirakan secara lokal dan semua komputasi ditangguhkan hingga evaluasi fungsi. Karena algoritma ini bergantung pada jarak untuk klasifikasi, menormalkan data pelatihan dapat meningkatkan akurasinya secara dramatis. Baik untuk klasifikasi dan regresi, teknik yang berguna dapat memberikan bobot pada kontribusi neighbors, sehingga neighbors yang lebih dekat berkontribusi lebih banyak pada rata-rata daripada yang lebih jauh. Sebagai contoh, skema pembobotan umum memberikan bobot 1/d kepada setiap neighbors, di mana d adalah jarak ke neighbors. Neighbors diambil dari sekumpulan objek yang kelasnya (untuk klasifikasi k-NN) atau nilai properti objek (untuk regresi k-NN) diketahui. Ini dapat dianggap sebagai training set untuk algoritma, meskipun tidak diperlukan langkah pelatihan eksplisit.

Dalam algoritma ini, nilai k yang terbaik itu tergantung pada jumlah data. Ukuran nilai k yang besar belum tentu menjadi nilai k yang terbaik begitupun juga sebaliknya. Langkah-langkah untuk menghitung algoritma k-NN:

1. Menentukan nilai k.
2. Menghitung kuadrat jarak euclid (query instance) masing-masing objek terhadap training data yang diberikan.
3. Kemudian mengurutkan objek- objek tersebut ke dalam kelompok yang mempunyai jarak

euclid terkecil.

4. Mengumpulkan label class Y
5. Dengan menggunakan kategori Nearest Neighborhood yang paling mayoritas maka dapat di prediksi nilai query instance yang telah di hitung.

Kelebihan K Nearest Neighbours yaitu mudah dipahami dan diimplementasikan, sangat non linear, dan asymptotically correct. Kekurangannya yaitu lambat saat prediksi, rentan terhadap perbedaan rentang variabel, rentan terhadap dimensionalitas yang tinggi, rentan terhadap variabel yang noninformatif, sensitif terhadap data pencilan, tidak menangani nilai hilang (missing value) secara implisit, sulit diinterpretasi

3. HASIL DAN PEMBAHASAN

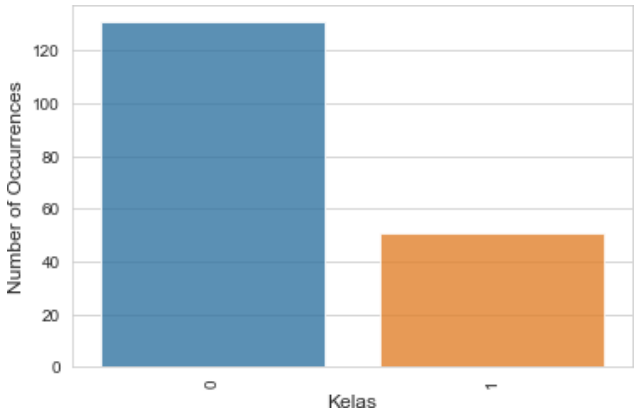
4.1 Data Vizualization

Terdapat 182 data yang berasal dari stasiun klimatologi bogor rentang tahun 2016-2017

```
Size of weather data frame is : (182, 8)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 182 entries, 0 to 181
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Stasiun      182 non-null    object
 1   Tanggal      182 non-null    object
 2   Tn           182 non-null    float64
 3   Tx           182 non-null    float64
 4   Tavg         182 non-null    float64
 5   RR           182 non-null    int64
 6   Hari_hujan   182 non-null    int64
 7   Besok_hujan  182 non-null    int64
dtypes: float64(3), int64(3), object(2)
```

Gambar 3. Data klimatologi

4.1.1 Untuk visualisasi data menggunakan barplot



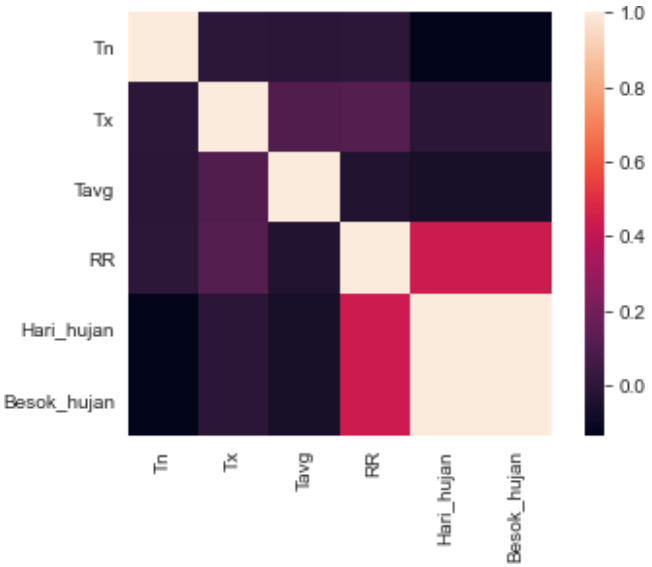
Gambar 4. Barplot

4.1.2 Visualisasi data dengan pairplot

Jurnal Riset Rumpun Ilmu Teknik (JURRITEK)
Vol.1, No.1 April 2022
e-ISSN: 2829-016X; p-ISSN: 2829-0178, Hal 88-97

Gambar 4. Pairplot

4.1.3 Visualisasi menggunakan plotting heatmap



Gambar 5. Plotting heat map

Warna dalam heatmap dapat menunjukkan frekuensi suatu peristiwa, kinerja berbagai metrik dalam kumpulan data, dan sebagainya. Palet warna di atas mewakili jumlah korelasi di antara variabel. Warna yang lebih terang menunjukkan korelasi yang tinggi. Untuk penelitian ini menggunakan ukuran sampel 30%, diasumsikan merupakan rasio ideal antara pelatihan dan pengujian.

```
Var correlation < 0.5%      Besok_hujan   Tn   Tx Tavg
Tx      0.002888  NaN  NaN  NaN
Data Final (182, 5)
X train shape:  (127, 4)
Y train shape:  (127,)
X test shape:  (55, 4)
Y test shape:  (55,)
```

Gambar 6. Data pelatihan

4.2 Classification

```
from sklearn.neighbors import KNeighborsClassifier

# We define the model
knncla = KNeighborsClassifier(n_neighbors=5,n_jobs=-1)

# We train model
knncla.fit(X_train, Y_train)

# We predict target values
Y_predict6 = knncla.predict(X_test)
```

```
# The confusion matrix
from sklearn.metrics import confusion_matrix
import seaborn as sns

knncla_cm = confusion_matrix(Y_test, Y_predict6)
f, ax = plt.subplots(figsize=(5,5))
sns.heatmap(knncla_cm, annot=True, linewidth=0.7, linecolor='cyan', fmt
='g', ax=ax, cmap="BuPu")
plt.title('KNN Classification Confusion Matrix')
plt.xlabel('Y predict')
plt.ylabel('Y test')
plt.show()
```

Gambar 7. Classification

4.3 Accurasy, Recall, Precision

```
#Accuracy
model1 = pd.DataFrame({
    'Model': ['KNN'],
    'Train Score': [train_acc_knncla],
    'Test Score': [test_acc_knncla]
})
model1.sort_values(by='Test Score', ascending=False)
```

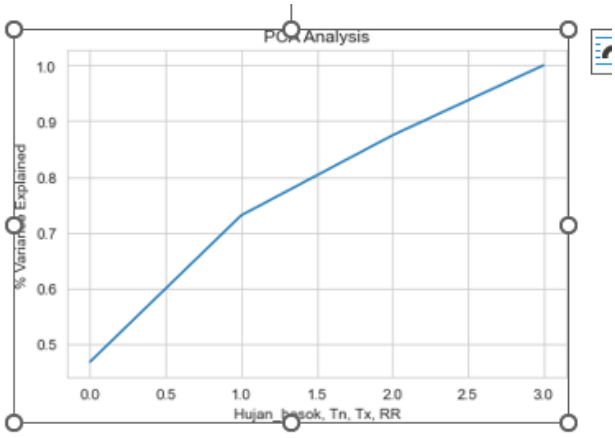
```
#Precision, Recall
from sklearn.metrics import average_precision_score
average_precision = average_precision_score(Y_test, Y_predict6)

print('Average precision-recall score: {0:0.2f}'.format(
      average_precision))
```

Gambar 8. Accuracy, recall dan precision score

Rata-rata skor precision-recall adalah sebesar 0.84

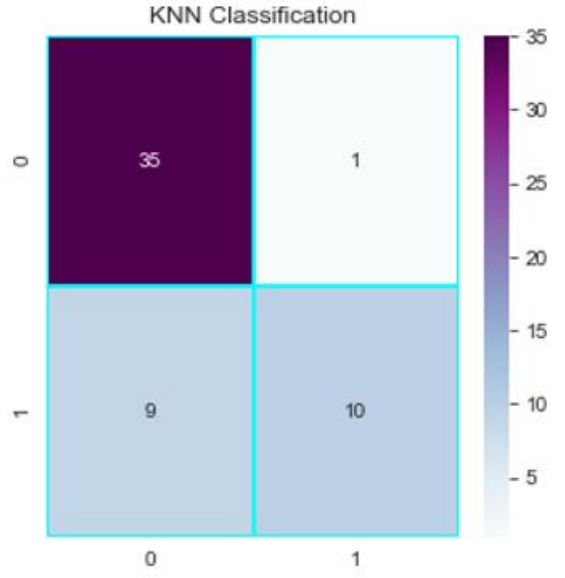
4.4 Featrues Selection

Principal Component Analysis (PCA) adalah teknik yang paling banyak digunakan dalam analisis data eksplorasi dan machine learning untuk model prediktif. Selain itu, PCA adalah teknik statistik tanpa pengawasan yang digunakan untuk memeriksa keterkaitan antara satu set variabel.



Gambar 9. Grafik PCA

```
# K-Nearest Neighbor classification
knncla.fit(X1_train, Y1_train)
Y1_predict6 = knncla.predict(X1_test)
knncla_cm = confusion_matrix(Y1_test, Y1_predict6)
score1_knncla= knncla.score(X1_test, Y1_test)
```




```
K-Nearest Neighbour Score    0.931408
Average precision-recall score: 0.79
```

Hasil akhir yang diperoleh untuk skor KNN sebesar 0.931408 dan rata-rata precision-recall sebesar 0.79.

4. KESIMPULAN DAN SARAN

Dengan memanfaatkan metode K- nearest Neighbour dengan teknik seleksi fitur PCA kita dapat memprediksi curah hujan di kota Bogor dengan baik, sehingga kita dapat mengetahui predeksi perkiraan curah hujan esok hari.

DAFTAR PUSTAKA

[1] F. Hermawan, H. Agung, "Implementasi Metode K-Nearest Neighbor Pada Aplikasi Data Penjualan PT. Multitek Mitra Sejati". *Jurnal Sains dan Teknologi*. ,Volume 4. Agustus 2017.

[2] Hasym, I E & Susilawati, I. "Klasifikasi Jenis Ikan Cupang Menggunakan Algoritma Principal Component Analysis (PCA) Dan K-Nearest Neighbors (KNN)". *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi.*

[3] Putri, Ayu Azlina. "Penerapan Data Mining Untuk Memprediksi Penjualan Buah Dan Sayur Menggunakan Metode K-Nearest Neighbor (Studi Kasus : PT. Central Brastagi Utama)". *RESOLUSI : Rekayasa Teknik Informatika dan Informasi*, Vol 1, No 6, Juli 2021.

[4] Wangge, M. "Penerapan Metode Principal Component Analysis (PCA) Terhadap Faktor-faktor yang Mempengaruhi Lamanya Penyelesaian Skripsi Mahasiswa Program Studi Pendidikan Matematika FKIP UNDANA:. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *5*(2), 974-988, 2021. https://doi.org/10.31004/cendekia.v5i2.465