Institut Riset dan Publikasi Indonesia (IRPI)
SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat
Homepage: https://journal.irpi.or.id/index.php/sentimas
Prosiding Seminar Nasional/ ISSN(E): 2830-3083
25 Agustus 2022/ Pages: 40-48

Comparison of Data Mining Methods for Predition of Floods with Naïve Bayes and KNN Algorithm

Perbandingan Metode Data Mining untuk Prediksi Banjir Dengan Algoritma *Naïve Bayes* dan KNN.

Cumel[1], David Zamri[2], Rahmaddeni[3], Syamsurizal[4]

[1,2,3,4]STMIK AMIK RIAU, Jl. Purwodadi Panam, Pekanbaru and 28294, Indonesia Indah Km 10

E-Mail: [1]2110031802123@sar.ac.id, [2]2110031802130@sar.ac.id, [3]rahmaddeni@sar.ac.id, [4]2110031802138@sar.ac.id

Abstract

Flooding is an event that needs to be watched out for because it is classified as a natural disaster. The Regional Disaster Management Agency (BPBD) is one of the government agencies in charge of conveying flood information in Jakarta. For floods, there are standards that will be achieved by BPBD, namely the status of floodgates at Pos Depok, Marina Ancol, Jembatan Merah, Katulampa, Flusing Ancol, Istiqlal and Manggarai. This flood dataset was taken from the flood database in DKI Jakarta from 1/1/2020 to 12/7/2020 which was taken from www.kaggle.com. To predict floods, data mining methods with classification capabilities are used. The approach used in "data mining" uses an activity process in the form of Knowledge Discovery in Databases (KDD), starting with the stages of selection, preprocessing, transformation, data mining and evaluation/interpretation. The techniques that will be used in this data mining classification model include four algorithms, namely k-Nearest Neighbors (kNN) and Naive Bayes. The classification method includes target variables and predictor variables. Predictors include Depok Post gates, Marina Ancol, Jembatan Merah, Katulampa, Flusing Ancol, Istiqlal and Manggarai. The software used in data processing is Rapid Miner software. The final result of the two algorithms is that the kNN algorithm is the best algorithm in flood prediction, accuracy value (88.94%), error (11.06%).

Keyword : *Data Mining, Floods, KNN, Naïve Bayes*

Abstrak

Banjir adalah suatu peristiwa yang perlu diwaspadai sebab tergolong musibah alam. Badan Penanggulangan Bencana Daerah (BPBD) merupakan salah satu lembaga pemerintahan yang bertugas menyampaikan informasi banjir di Jakarta. Untuk banjir terdapat standar yang akan dicapai oleh BPBD yaitu status pintu air di Pos Depok, Marina Ancol, Jembatan Merah, Katulampa, Flusing Ancol, Istiqlal dan Manggarai. *Dataset* banjir ini diambil dari database banjir di DKI Jakarta mulai tanggal 1/1/2020 sampai tanggal 12/7/2020 yang diambil dari www.kaggle.com. Untuk memprediksi banjir, digunakan metode data mining dengan kemampuan klasifikasi. Pendekatan yang digunakan dalam "data mining" ini memakai proses kegiatan berupa Knowledge Discovery in Databases (KDD), diawali dengan tahapan seleksi, *preprocessing*, transformasi, data mining dan evaluasi/interpretasi. Teknik yang akan dipakai dalam model klasifikasi data mining ini meliputi empat algoritma yaitu *k-Nearest Neighbors* (kNN) dan *Naive Bayes*. Metode klasifikasi meliputi variabel target dan variabel prediktor. Prediktornya antara lain gerbang Pos Depok, Marina Ancol, Jembatan Merah, Katulampa,Flusing Ancol, Istiqlal dan Manggarai. Perangkat lunak yang digunakan dalam pengolahan data adalah perangkat lunak Rapid Miner. Hasil akhir dari kedua algoritma tersebut adalah algoritma kNN merupakan algoritma yang paling bagus dalam prediksi banjir, nilai akurasi (88,94%), error (11,06%).

Kata Kunci: Banjir, *Data Mining*, KNN, *Naïve Bayes*

1. PENDAHULUAN

Banjir merupakan suatu musibah alam yang diprediksi oleh Badan Penanggulangan Bencana Daerah (BPBD). *Dataset* banjir ini diambil dari database banjir di DKI Jakarta mulai tanggal 1/1/2020 sampai tanggal 12/7/2020. Menurut data banjir di database, ada sebagian prediktor yang bisa dipakai guna memperkirakan variabel target yaitu status pintu air di Katulampa, Depok Post, Maggarai, Istiqlal, Jembatan Merah, Flusing Ancol, dan Marina Ancol. Salah satu metode untuk melakukan prediksi adalah dengan memakai data mining, yaitu akan mencari model yang ada di database banjir guna mengklasifikasikan banjir yang terjadi. Metode yang akan dipakai tergolong ke dalam fungsi klasifikasi dengan berbagai algoritma.

Data Mining adalah teknik guna mengekstraksi atau "menambang" informasi dari kumpulan data yang besar. Data mining adalah meninjau berbagai informasi guna mendapat korelasi yang tak terduga dan merangkum informasi melalui metode yang berbeda agar bisa dimengerti dan berguna untuk pemilik informasi [1].

Klasifikasi memiliki berbagai metode diantaranya *K-Nearest Neighbor m, Nave Bayes Classifier, Artificial Neural Network, Classification Tree, Support Vector Machine, Discriminant Analysis*, dan sebagainya. Dalam riset ini metode klasifikasi yang dipakai yaitu dengan membandingkan nilai Accuracy dan Error antara *KNN dan Naïve Bayes Classifier*. Ukuran rasio prediksi yang benar dengan jumlah sampel yang dinilai yaitu "akurasi" sedangkan untuk yang salah yaitu "kesalahan".

Beberapa penelitian mengenai penerapan bagaia riset terkait implementasi data mining guna mencari data yang ada di database mahasiswa diantaranya adalah Muhammad Firdaus, Rahmaddeni, Yustis Maruhawa mengenai perbandingan metode data mining untuk prediksi curah hujan dengan algoritma C4.5, *Naïve Bayes*, dan kNN yang menghasilkan algoritma C4.5 merupakan algoritma terbaik untuk memprediksi curah hujan dengan nilai *accuracy (*88,03%), *error* (11,97%). Sri Widaningsih melakukan analisa perbandingan metode data mining guna prediksi nilai dan waktu kelulusan mahasiswa dengan algoritma *Naïve Bayes*, C.45, knn, dan svm. Diperoleh *Naïve bayes* memberikan hasil yang terbaik [2].

2. METODE PENELITIAN

2.1. Sumber Data

Data yang dipakai pada riset yaitu data sekunder yang didapat dari BPBD DKI Jakarta mulai tanggal 1/1/2020 sampai tanggal 12/7/2020 dan bisa di cari memalui website www.kaggle.com.

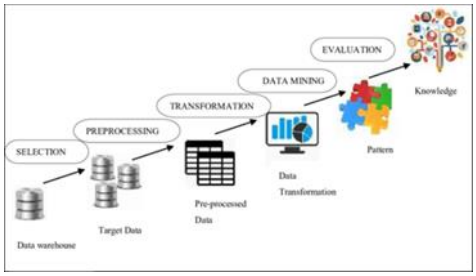
2.2. Variabel Penelitian

Variabel yang dipakai pada riset ini meliputi variabel terikat dan bebas.

1. Variabel terikat (Y) pada riset ini adalah status banjir yang dibagi menjadi dua jenis yaitu tergenang dan tidak tergenang.
2. Variabel bebas (X) yang dipakai ada empat variabel yaitu status gerbang Pos Depok, Marina Ancol, Jembatan Merah, Katulampa, Flusing Ancol, Istiqlal dan Manggarai.

2.3. Tahapan Analisis Data

Model yang dipakai pada riset ini mencakup berbagai alur dalam proses *knowledge discovery* (KDD) dalam *database*. Gambar 6 merupakan ilustrasi dari setiap tahapan yang akan dilakukan. Proses KDD diawali dengan penetapan visi dan berakhir dengan evaluasi [3]. Alur KDD bisa diketahui pada Gambar 1 dan Gambar 2.



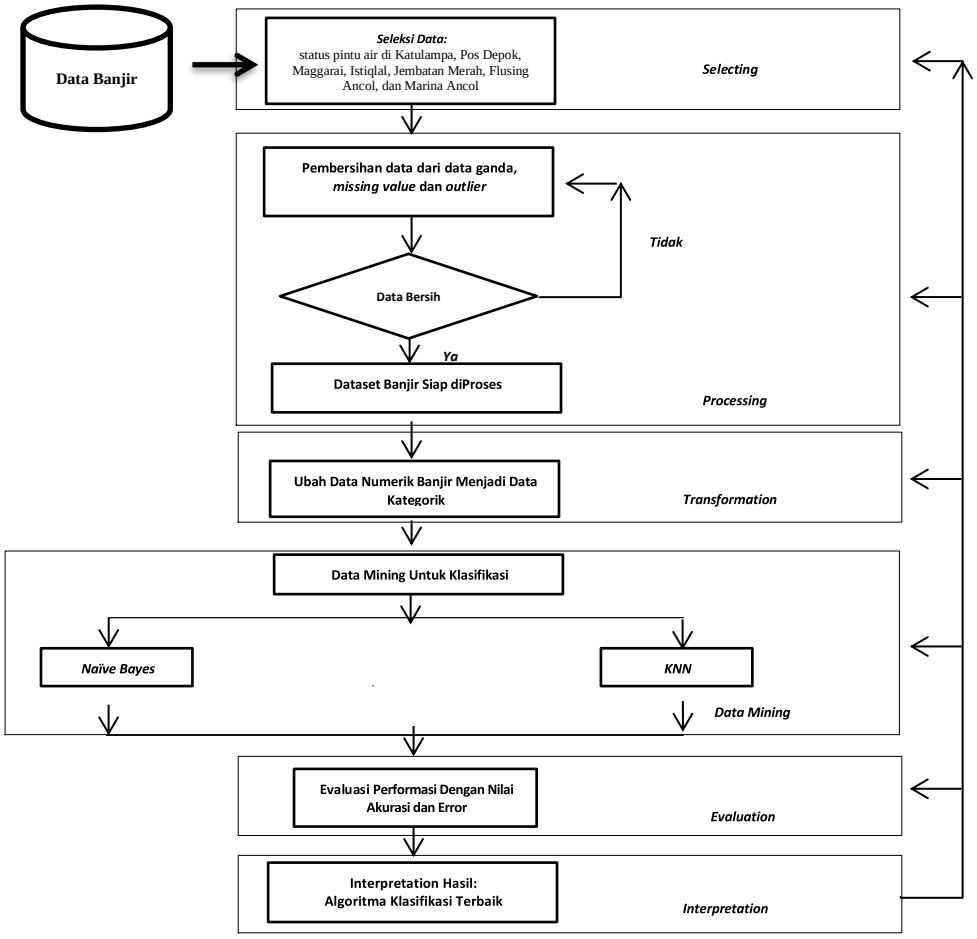
Gambar 1. Tahapan proses KDD

Uraian Gambar 1 dan Gambar 2.

1. *Selection*

   Dalam alur ini, pemilihan data cuaca meliputi variabel target dan prediktor. Prediktornya antara lain status gerbang Katulampa, Pos Depok, Maggarai, Istiqlal, Jembatan Merah, Flusing Ancol, dan Marina Ancol. Sedangkan variable target adalah Banjir.

2. *Preprocessing*
   Data yang diperoleh sama dengan jumlah lulusan program penelitian teknik informasi. Menurut data yang diperoleh, jika ada data yang hilang, data duplikat atau outlier, lakukan pembersihan data.
3. *Transformation*
   Selepas data dibersihkan dari kesalahan, data ditransformasikan menurut tipe data pada tahap transformasi, dan tipe data diklasifikasikan sebagai data kategorikal. Tabel 2 di bawah ini adalah kelompok variabel target dan prediktor.



Gambar 2. Tahapan penelitian berdasarkan KDD

Tabel 1. Kategori Variabel Prediktor dan Variabel Target

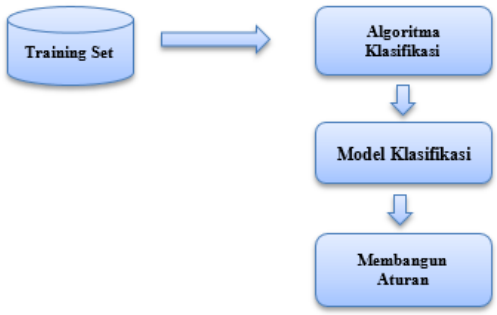
Variabel Target	Kategori
Banjir	Tidak Banjir Banjir
Variabel Prediktor	
Katulampa	0 – 1000 m^3/s
Pos Depok	0 – 1000 m^3/s
Manggarai	0 – 1000 m^3/s
Istiqlal	0 – 1000 m^3/s
Jembatan Merah	0 – 1000 m^3/s
Flusing Ancol	0 – 1000 m^3/s
Marina Ancol	0 – 1000 m^3/s

4. *Data Mining*
   Dalam alur ini, teknik data mining yang tepat dipilih. Guna fungsi klasifikasi, dipakai *Naive Bayes* dan algoritma kNN. Sebab klasifikasi adalah supervised learning, beserta alur alur pada model *supervised learning* [1].
5. *Evaluation*
   Fase ini dipakai guna menilai prediksi yang dikeluarkan oleh keempat algoritma dan metode algoritma yang dipilih yang melahirkan nilai yang hammendekati klasifikasi data aslinya. Evaluasi dilaksanakan dengan memakai metode confusion *matrix* dan kurva *Receiver Operating Characteristic* (ROC). Nilai performa yang dipakai adalah error dan akurasi.

2.4. Metode Analisis Data

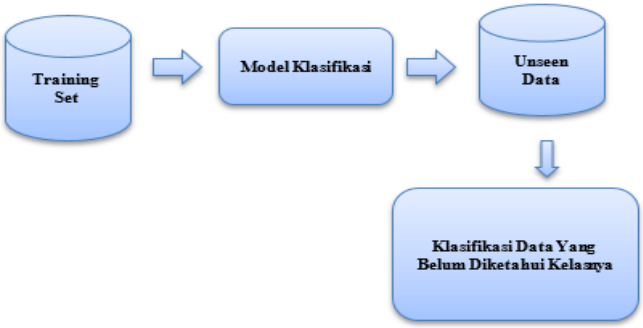
Algoritma data *mining* bisa dibagi menjadi tiga [4], yaitu *supervised unsupervised*, dan *semi-supervised*. Pada *supervised learning*, algoritma bekerja pada bebrapa kategori data yang diberi label atau diketahui. Dalam proses supervised learning, data belum dikenal dengan label atau kategori, dan dipakai proses algoritmik agar dapat mengelompokkan data menurut data terdekat. Sedangkan pada pembelajaran semi-supervised, sebagian kecil data sudah diberi label, dan sebagian data belum. Klasifikasi termasuk dalam pembelajaran terawasi. Proses klasifikasi dibagi menjadi dua tahap [5] yaitu :

1. Tahap membangun model
   Dalam tahap ini, model klasifikasi dirancang dari data yang ditetapkan kelas. Data sampel yang dipakai disebut data pembelajaran (*training set*). Alur ini disebut proses induksi dan diperlihatkan pada Gambar 3.



Gambar 3. Tahapan Membangun Model

2. Tahap memakai model klasifikasi
   Dalam alur ini, model dipakaipada data dengan kategori yang tidak diketahui. Proses implementasi model klasifikasi guna memprediksi label kelas dari data pada himpunan dengan memakai data uji (*test set*), proses ini disebut inferensi. Lihat Gambar. 4.



Gambar 4. Tahapan Memakai Model

2.5. Algoritma Klasifikasi

Pada klasifikasi, variabel target kategoris diklasifikasikan ke dalam kategori yang sudah ditetapkan, seperti kategori pelanggan bermasalah atau tidak bermasalah, hewan yang termasuk pada kelompok mamalia, ikan, reptil, amfibi atau burung. Setiap *record* data mining berisikan informasi terkait variabel target dan satu set input atau variabel prediktor. Tiap tiap algoritma klasifikasi yang dipakai menciptakan model yang

terbaik, mengkaitkan data input dan kategori klasifikasi yang sudah pernah ditetapkan. Tiap tiap algoritma dapat menciptakan klasifikasi yang beragam. Algoritma terbagus bisa diketahui melalui seberapa akurat model mengklasifikasikan data versus data aktual. Di bawah ini adalah penjelasan dari *Naive Bayes* dan kNN.

1. *Naïve Bayes*

Klasifikasi *Bayes* adalah pengelompokan statistik yang memperkirakan probabilitas keanggotaan kelas, seperti probabilitas bahwa sebuah tuple milik kelas tertentu. Penggolongan ini didasarkan pada teorema Bayes. Di Naive Bayes, peran nilai atribut di kelas tertentu tidak tergantung pada nilai atribut lainnya**.** Rumus *Naive Bayes* [6] . Rumus algoritma *naive Bayes* diperlihatkan dalam persamaan (3) dibawah ini:

$$P(Y \mid X) = P(Y)||P(X \mid Y) \tag{3}$$

Dimana :

P(X | Y) : probabilitas data dengan vektor X pada kelas Y

P(Y) : probabilitas awal kelas Y dan P(Xi | Y) adalah probabiltas independen kelas Y pada semua fitur dalam vektor X

2. *k Nearest Neighbor*

Algoritma adalah cara untuk menemukan kasus dengan menghitung kedekatan antara kasus baru dan kasus lama. Disebut juga pembelajar malas sebab hanya meninjau keakraban dengan tetangga (*neighbor*). Pada Persamaan (4), salah satu rumus jarak yang dipakai pada K tetangga terdekat adalah jarak *Euclidean* [7]:

$$^{d}Euclidian\ (x,y) = \sqrt{\sum i(xi - i)^2} \tag{4}$$

2.6. Alat Evaluasi

Klasifikasi biner adalah model statistik dan komputasi yang membagi kumpulan data menjadi dua jenis, positif dan negatif. Tabel 2 di bawah ini adalah matriks konfusi yang menjelaskan metrik kinerja klasifikasi.

Tabel 2. *Confusion Matrix*

	Prediksi	
Aktual	*Positive*	*Negative*
Positive	*True Positive*	*False Negative*
Negative	*False Positive*	*True Negative*

Indikator kinerja tergolong pada tahap evaluasi. Berbagai metrik kinerja guna teknik klasifikasi adalah akurasi, kesalahan, dan area karakteristik operasi penerima ROC di bawah kurva AUC. Akurasi adalah ukuran rasio prediksi yang benar dengan jumlah sampel yang dinilai. Kesalahan adalah ukuran rasio prediksi yang salah dengan jumlah total sampel yang dinilai.

Area Under the Curve (AUC) adalah ukuran perbedaan kinerja. Pada pengelompokan data mining, nilai AUC bisa dibagi menjadi berbagai kelompok [7].

a. 0.90-1.00 = Klasifikasi sangat baik
b. 0.80-0.90 = Klasifikasi baik
c. 0.70-0.80 = Klasifikasi cukup
d. 0.60-0.70 = Klasifikasi buruk
e. 0.50-0.60 = Klasifikasi salah

$$Akurasi = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

$$Error = \frac{FN + FP}{TP + TN + FN + FP} \tag{6}$$

RapidMiner

Pada pengolahan data mining, perangkat lunak sering dipakai sebagai alat bantu. Berbagai perangkat lunak penambangan data termasuk RapidMiner, Weka, Clementine, Tanagra, dll. Berdasarkan www.rapidminer.com, perangkat lunak rapidminer bertindak sebagai perancang proses visual untuk menganalisis ilmu data dan pembelajaran mesin pada tim dimulai dari analis sampai pakar. Perangkat lunak

ini mudah digunakan dan bias menghimpun data dari semua sumber seperti database, media sosial, cloud, dokumen atau aplikasi bisnis. Selain itu, memungkinkan eksplorasi statistik dan visualisasi data. Berbagai model pembelajaran mesin dan model validasi tersedia.

2.7. Penelitian Terdahulu

Berbagai riset tentang implementasi data mining guna mencari informasi yang ada dalam database mahasiswa diantaranya adalah Muhammad Firdaus, Rahmaddeni, Yustis Maruhawa mengenai perbandingan metode data mining untuk prediksi curah hujan dengan algoritma C4.5, Naïve Bayes, dan kNN yang menghasilkan algoritma C4.5 merupakan algoritma terbaik untuk memprediksi curah hujan dengan nilai *accuracy (*88,03%), *error* (11,97%).

Sri Widaningsih melakukan analisa perbandingan metode data mining guna perkiraan nilai dan waktu kelulusan mahasiswa dengan algoritma Naïve Bayes, C.45, knn, dan svm. Diperoleh Naïve bayes memberikan hasil yang terbaik.

Elisabet Sinta Romaito, M. Khairul Anam, Rahmaddeni , dan Aniq Noviciate Ulfah (2021) melakukan perbandingan algoritma svm dan nbc pada analisa sentimen pilkada pada twitter dengan hasil Accuracy 81.7 memiliki recall sebesar 81.7 dan presisi 80% adalah hasil NBC, sementara Accuracy 80.7 memiliki recall sebesar 80,7 dan presisi 84% adalah hasil SVM. Jadi bisa diambil dari di atas bahwa algoritma NBC lebih baik dari Precision pada hal akurasi dan recall, dan Precision lebih baik daripada algoritma SVM.

Rahmaddeni, M. Khairul Anam, Yuda Irawan, Susanti, Jamaris (2021). Perbandingan Support Vector Machine dan XGBSVM Dalam Menganalisis Opini Publik Vaksinasi Covid-19 didapat kita ketahui SVM mendapatkan akurasi tertinggi yakni 83% dengan spliting data 90:10, kemudian XGBSVM menghasilkan akurasi 79% dengan spliting data 90:10.

2.8. Analisa dan Perancangan

1. *Data Selection*

Sumber data mentah yang dipakai pada riset ini adalah Dataset banjir di DKI Jakarta mulai tanggal 1/1/2020 sampai tanggal 12/7/2020. Data nilai berasal dari alamat *website* www.kaggle.com.

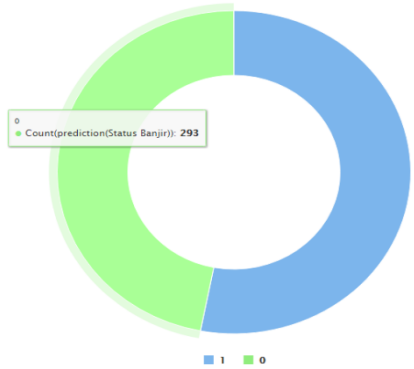
2. *Preprocessing Data*

Guna melihat curah hujan setiap harinya dilaksanakan perhitungan kembali dari data suhu, kelembaban, lama penyinaran dan kecepatan angin sebab ada berbagai redudansi data nilai maksimum dan minimum karena faktor cuaca yang seringkali berubah-ubah. Contoh dari pengumpulan data disusun pada Tabel 3.

Tabel 3. Data awal Banjir

Katulam...	Pos Dep...	Manggar...	Istiqlal	Jembata...	Flusing ...	Marina A...	Status B...
47.000	167.000	639.000	206.000	164.000	180.000	159.000	0.000
44.000	75.000	680.000	215.000	175.000	181.000	114.000	0.000
42.000	159.000	735.000	187.000	192.000	197.000	169.000	0.000
41.000	161.000	845.000	238.000	223.000	203.000	178.000	0.000
40.000	154.000	889.000	280.000	262.000	223.000	191.000	1.000
39.000	153.000	927.000	288.000	267.000	219.000	190.000	1.000
40.000	202.000	934.000	293.000	269.000	218.000	196.000	1.000
77.000	207.000	931.000	295.000	271.000	221.000	199.000	1.000
130.000	234.000	934.000	297.000	277.000	231.000	207.000	1.000
141.000	287.000	927.000	297.000	278.000	237.000	211.000	1.000
139.000	313.000	911.000	302.000	279.000	235.000	211.000	1.000
116.000	410.000	901.000	302.000	277.000	234.000	208.000	1.000

Berdasarkan data cuaca pada BPBD DKI Jakarta mulai tanggal 1/1/2020 sampai tanggal 12/7/2020 menciptakan klasifikasi "Tidak Banjir (0)" atau "Banjir (1)" bisa diketahui pada Gambar 5.



Gambar 5. Perbandingan klasifikasi kelas "Tidak Banjir (0)" atau "Banjir (1)"

3. *Transformation*

Dataset yang akan diolah dituang ke dalam tabel data awal kemudian diubah menjadi beberapa jenis data numerik yaitu banjir. Bentuk dataset yang ditransformasi ditunjukkan pada Tabel 4.

Tabel 4. Data yang Telah Ditransformasi

Row No.	Status Banjir	prediction...	confidence(0)	confidence(1)	Katula...	Pos Dep...	Mang...	Istiqlal	Jemba...	Flusin...	M...
1	1	0	0.588	0.412	39	153	927	288	267	219	190
2	1	0	0.567	0.433	71	233	892	296	264	209	177
3	0	0	1	0	55	173	962	334	284	209	177
4	0	0	1	0	51	164	919	330	282	214	193
5	0	0	1.000	0	46	151	774	301	268	196	185
6	1	0	0.627	0.373	39	129	672	217	195	?	170
7	0	0	1.000	0	40	121	779	235	222	363	211
8	0	1	0	1	40	122	475	207	193	336	185
9	0	1	0.200	0.800	72	163	475	184	194	297	198
10	0	0	1	0	81	183	729	173	170	250	158
11	1	1	0.196	0.804	61	168	845	249	233	270	180
12	0	1	0.377	0.623	63	151	892	242	250	280	187
13	0	0	0.619	0.381	63	156	866	205	214	281	170
14	0	0	0.623	0.377	61	130	842	182	190	291	158
15	0	0	0.607	0.393	60	115	832	169	178	317	157

ExampleSet (624 examples, 4 special attributes, 7 regular attributes)

4. *Data Mining*

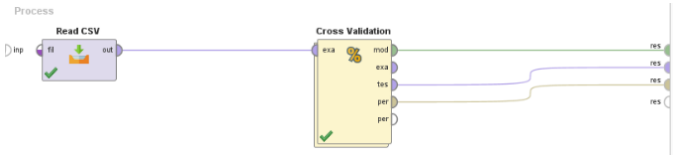
Dalam pengolahan data, tahap pemodelan dari proses pengelompokkan dilakukan dengan menetapkan algoritma *Naive Bayes* dan kNN. Pengolahan data dilakukan dengan menggunakan *software* RapidMiner 9.10.008. Jumlah data yang akan diolah sejumlah 624 data. Proses pengambilan data Naive Bayes dan kNN di Rapidminer diekstraksi langsung dari data hasil transformasi dalam format excel, seperti terlihat pada Tabel 4 di atas. Selain itu, validasi silang dilakukan pada data yang diperoleh. Teknik validasi yang dipakai dalam proses klasifikasi dalam penelitian ini adalah *k-Fold Cross Validation.*

K-fold cross-validation adalah metode statistik guna menilai kinerja model klasifikasi, di mana data dibagi menjadi dua bagian, data proses pelatihan dan data uji. Validasi silang K-fold dipakai sebab menurunkan waktu komputasi dengan terus mempertahankan nilai akurasi yang diperkirakan. Nilai k dikalikan 10 alhasil dari 624 data terdapat 10 himpunan bagian data yang berukuran sama, yaitu sekitar 62,4 atau 63 data. Dalam 10 himpunan bagian, 561 data menjadi data pelatihan dan 63 data menjadi data uji.

3. HASIL DAN PEMBAHASAN

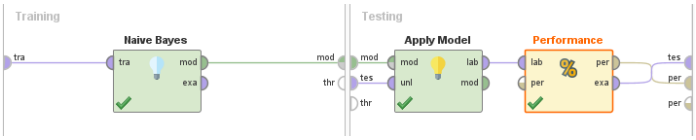
3.1. Hasil Perhitungan Klasifikasi dengan RapidMiner

Proses pengumpulan data dilakukan sesuai dengan algoritma data mining yang dipakai dalam riset ini yaitu Naive Bayes dan kNN. seperti Gambar 6.



Gambar 6. Proses Pengambilan Data Guna Algoritma *Naïve Bayes* dan kNN.

Pada data pelatihan, algoritma Naive Bayes diterapkan pada teknik klasifikasi, seperti yang ditunjukkan pada Gambar 7.



Gambar 7. Model Klasifikasi dengan Algoritma Naïve Bayes

Keluaran dari hasil performansi algoritma adalah klasifikasi siswa yang tergolong pada pengelompokkan kelas "tidak banjir" atau "banjir" yang dijalankan pada data uji. Jumlah data yang diperkirakan dengan benar oleh algoritma *Naive Bayes* diperlihatkan pada Tabel 5 *Confusion Matrix*.

Tabel 5. *Confusion Matrix* Algoritma Naïve Bayes

accuracy: 74.38% +/- 6.61% (micro average: 74.36%)

	true 0	true 1	class precision
pred. 0	202	69	74.54%
pred. 1	91	262	74.22%
class recall	68.94%	79.15%	

Keterangan tabel 5 adalah:

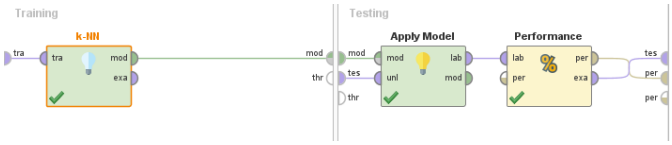
a. Jumlah data sebenarnya TIDAK BANJIR dan diperkirakan TIDAK BANJIR adalah 202.
b. Jumlah data sebenarnya BANJIR dan diperkirakan BANJIR adalah 262.
c. Jumlah data sebenarnya TIDAK BANJIR dan diperkirakan BANJIR adalah 91.
d. Jumlah data sebenarnya BANJIR dan diperkirakan TIDAK BANJIR adalah 69.

Evaluasi bagi model ini memakai nilai akurasi, *error*.
Akurasi dari model yaitu :

$$\text{Akurasi} = \frac{202 + 262}{202 + 262 + 91 + 69} = 74{,}36\%$$

$$\text{Error} = \frac{91 + 69}{202 + 262 + 91 + 69} = 25{,}64\%$$

Pada data latih, algoritma kNN diterapkan pada teknik klasifikasi, seperti terlihat pada Gambar 8.



Gambar 8. Model Klasifikasi dengan Algoritma kNN

Jumlah data yang diperkirakan dengan benar oleh algoritma kNN diperlihatkan pada matriks konfusi seperti yang diperlihatkan pada Tabel 6.

Tabel 6. *Confusion Matrix* Algoritama kNN dengan k = 5

accuracy: 88.93% +/- 3.01% (micro average: 88.94%)

	true 0	true 1	class precision
pred. 0	265	41	86.60%
pred. 1	28	290	91.19%
class recall	90.44%	87.61%	

Keterangan tabel 6 adalah:

a. Jumlah data sebenarnya TIDAK BANJIR dan diprediksi TIDAK BANJIR adalah 265.
b. Jumlah data sebenarnya BANJIR dan diprediksi BANJIR adalah 290.
c. Jumlah data sebenarnya TIDAK BANJIR dan diprediksi BANJIR adalah 28.
d. Jumlah data sebenarnya BANJIR dan diprediksi TIDAK BANJIR adalah 41.

Evaluasi bagi model ini memakai nilai akurasi, *error*.
Akurasi dari model yaitu:

$$\text{Akurasi} = \frac{265 + 290}{265 + 290 + 28 + 41} = 88{,}94\%$$

$$\text{Error} = \frac{28 + 41}{265 + 290 + 28 + 41} = 11{,}06\%$$

Hasil performansi dari masing-masing model yaitu akurasi dan error kemudian akan dibandingkan guna melihat algoritma mana yang lebih baik saat memperkirakan curah hujan yang terjadi. Perbandingan antara 2 algoritma dapat dilihat pada tabel 7.

Tabel 7. Komparasi Nilai Performansi Setiap Algoritma

Algoritma	*Accuracy*	*Error*
kNN, k = 5	88,94%	11,06%
Naïve Bayes	74,36%	25,64%

Dari hasil perbandingan tersebut dapat diketahui bahwa algoritma kNN memiliki nilai yang paling baik dibandingkan dengan algoritma lainnya pada semua kategori performansi. Untuk nilai akurasi "baik", itu adalah nilai minimum untuk kesalahan. Nilai Naive Bayes termasuk dalam kategori "cukup".

4. KESIMPULAN

Berdasar pada pembahasan bisa diambil kesimpulan:

1. Dengan menggunakan teknik data mining, akurasi dan kesalahan bisa didapat dalam database curah hujan.
2. Menurut algoritma-algoritma yang diuji seluruhnya bias dipakai guna memperkirakan curah hujan, dapat dilihat dari nilai *accuracy* dan *error* dari seluruh algoritma yang ada tergolong “baik” dan “sedang” dan “cukup”
3. Menurut hasil evaluasi, dibandingkan dengan algoritma Naive Bayes, algoritma kNN mempunyai nilai akurasi tertinggi dan error terkecil, serta merupakan algoritma terbaik guna memperkirakan tingkat kelulusan yang diinginkan.

REFERENSI

[1] Larose. 2005. *Discovering Knowledge in Data*.Canada: Wiley-Interscience

[2] Sri Widaningsih (2019). Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4.5, Naïve Bayes, Knn, Dan Svm. Jurnal Tekno Insentif | ISSN (p): 1907-4964 | ISSN (e): 2655-089X

[3] Tomar, D., & Agarwal, S. (2013). *A survey on Data Mining approaches for Healthcare*. International Journal of Bio- Science and Bio -Technology, 5 , 241-266

[4] Neelamegam, S., & Ramaraj, E. (2013). *Classification algorithm in Data mining*: An Overview. International Journal of P2P Network Trends and Technology (IJPTT), 3, 1-5.

[5] Annasaheb, A.B., & Verma, V.K. (2016). *Classification Techniques*: A Recent Survey. International Journal of Emerging Technologies in Engineering Research (IJETER), 4, 51-54.

[6] Suyatno. (2017) Data Mining Untuk Klasifikasi dan Klasterisasi Data. Bandung: Informatika

[7] Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques*. Berlin: Springer