

Analisis Hubungan antar Faktor dan Komparasi Algoritma Klasifikasi pada Penentuan Penundaan Penerbangan

Danny Ibrahim^{1*)}

¹Program Magister Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
email: ¹dannyibm10@gmail.com

Abstrak > Penyebab terjadinya kecelakaan pesawat diantaranya karena faktor cuaca sehingga diperlukan suatu system yang dapat membantu menentukan penundaan penerbangan karena kondisi cuaca dengan tepat. Penentuan penundaan penerbangan yang digunakan dalam penelitian ini adalah metode Proses Standard pada Data Mining (CRISP-DM) dengan algoritma *Correlation matrix* (mirip Asosiasi), *Klasifikasi*, dan *Feature Selection* yang dilakukan dengan langkah-langkah sebagai berikut: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Pada *Klasifikasi* dibandingkan 5 Algoritma yaitu *Decision Tree (C4.5)*, *Naive Bayes (NB)*, *K-Nearest Neighbor (K-NN)*, *Random Forest (RF)*, dan *Logistic Resgion (LogR)*. Sedangkan pada *Feature Selection* dibandingkan 3 metode yaitu *Filter Selection*, *Forward Selection (FS)*, dan *Backward Elimination (BE)*. Kesimpulan hasil penelitian adalah: metode CRISP-DM dapat digunakan untuk rekomendasi dalam menentukan penundaan penerbangan dengan baik. Metode CRISP-DM yang digunakan menerapkan beberapa peran data mining yaitu Korelasi, *Klasifikasi*, dan *Featrure Selection*. Dari masing-masing peran tersebut diketahui bahwa faktor yang paling berpengaruh adalah tekanan udara sedangkan faktor yang lain korelasinya sangat kecil, kemudian model terbaiknya adalah *Naive Bayes (NB)* untuk *Klasifikasi* dan *Naive Bayes + Backward Elimination (NB + BE)* untuk *Feature Selection*.

Kata Kunci > *Data Mining*, *CRISP-DM*, *Klasifikasi*, *Correlation Matrix*, *Feature Selection*, *Penundaan Penerbangan*, *Cuaca*, *Naive Bayes*, *Backward Elimination*.

I. PENDAHULUAN

Tak dapat dipungkiri, bahwa perkembangan data dewasa ini berkembang dengan sangat pesat, baik itu perkembangan jumlah jenis datanya maupun perkembangan jumlah pemakainya. Data mining berperan dalam penggalian informasi untuk mendapatkan pengetahuan yang terkandung dalam data tersebut [1].

Peningkatan kebutuhan terhadap analisa dan pengolahan data yang demikian pesat ini harus pula didukung oleh suatu metode yang dapat mengambil pengetahuan dari data tersebut. Metode CRISP-DM termasuk dalam kategori metode *Data Mining* yang bertujuan untuk memperoleh suatu pola / pengetahuan dari data yang dimanfaatkan untuk menyelesaikan suatu masalah [2].

Salah satu cara untuk menyelesaikan masalah adalah

*) penulis korespondensi

dengan mengolah data sehingga informasi yang terkandung didalamnya dapat diambil untuk mendapatkan pola/pengetahuan yang diperlukan untuk menyelesaikan masalah tersebut. Dalam hal ini peran *Data Mining* digunakan untuk mengatasi masalah dimana seringkali terjadinya kecelakaan pesawat disebabkan karena kondisi cuaca yang buruk sehingga diperlukan suatu metode yang dapat digunakan untuk membantu dalam menentukan penundaan penerbangan yang salah satunya adalah dengan menggunakan metode CRISP-DM yang bertujuan untuk mendapatkan pola/model pengetahuan dari data yang diperoleh untuk rekomendasi dalam pengambilan keputusan dalam membantu menentukan penundaan penerbangan sehingga dengan pengambilan keputusan yang tepat, volume terjadinya kecelakaan pesawat dapat berkurang dengan signifikan.

Dalam penentuan penundaan penerbangan ini digunakan metode klasifikasi dengan menggunakan 5 algoritma yaitu: *Decision Tree (C4.5)*, *Naive Bayes (NB)*, *K-Nearest Neighbor (KNN)*, *Random Forest (RF)*, dan *Logistic Regression (LogR)*. Dalam klasifikasi atribut dataset dapat berupa numerik atau nominal dan labelnya berupa nominal.

Tujuan dari penelitian ini adalah mencari hubungan antar faktor pada penundaan penerbangan. Melakukan perbandingan 5 algoritma klasifikasi untuk menentukan penundaan penerbangan dan melakukan perbandingan algoritma *Feature Selection* pada algoritma terbaik untuk menentukan penundaan penerbangan. Dengan demikian dapat membantu dalam pengambilan keputusan penundaan penerbangan yang yang tepat sehingga dapat mengurangi terjadinya kecelakaan pesawat.

II. TINJAUAN STUDI

A. Penelitian Terkait

Beberapa penelitian terdahulu yang dijadikan rujukan terkait dengan topik antara lain: Penelitian perancangan Sistem Prediksi Cuaca Berbasis *Logika Fuzzy* Untuk Kebutuhan Penerbangan Di Bandara Juanda – Surabaya [3]. Prediksi Cuaca Berbasis *Logika Fuzzy* untuk Rekomendasi Penerbangan di Bandar Udara Raja Haji Fisabilillah pada tahun 2011 [4]. Penelitian lain dilakukan untuk perbaikan Metode Prakiraan Cuaca Bandara Abdulrahman Saleh dengan Algoritma *Neural Network Backpropagation* pada tahun 2013 [5], dan Implementasi Algoritma C 4.5 untuk Menentukan Tingkat Bahaya Tsunami pada tahun 2011 [6]. Penerapan Algoritma Genetika Untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan Menggunakan

Naive Bayes pada tahun 2016 [7] dengan Akurasi mencapai 89.50 % dan AUC sebesar 0.919, ini berarti termasuk dalam kategori *Excellent Classification*.

B. Data Mining

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [8]. Pengelompokan Data Mining menurut Larose, data mining dibagi menjadi beberapa kelompok berdasarkan peran yang dapat dilakukan [9], yaitu:

- Estimasi*: Estimasi atributnya numerik, variabel label estimasi juga numerik.
- Prediksi*: Prediksi hampir sama dengan estimasi, kecuali bahwa dalam prediksi ada time series dan nilai dari hasil akan ada di masa mendatang.
- Klasifikasi*: Dalam klasifikasi, terdapat label variabel kategori.
- Pengklusteran*: *Clustering* merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik (*similarity*) antara satu data dengan data yang lain. *Clustering* merupakan salah satu metode *data mining* yang bersifat tanpa target/label (*unsupervised*).
- Asosiasi*: Metode asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam suatu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

C. Feature Selection

Metode *Feature Selection* dapat digunakan untuk mengurangi dimensi data dari atribut yang berlebihan misalnya ada data yang terduplikasi, kemudian atribut yang tidak relevan yang mengandung informasi yang tidak berguna untuk peran data mining, misalnya ID pelajar tidak berpengaruh pada prediksi GPA pelajar.

Feature Selection diklasifikasikan menjadi tiga yaitu: *Wrapper*, *Filter* dan *Hybrid*.

- Dalam pendekatan *Filter*, analisis statistik dari set fitur diperlukan tanpa menggunakan beberapa model pembelajaran.
- Dalam pendekatan *Wrapper*, diasumsikan model pembelajaran yang telah ditentukan dimana fitur yang terpilih yang menentukan performa model pembelajaran tertentu. Pendekatan *Wrapper* ini terdiri dari *Forward Selection* dan *Backward Elimination*.
- Pendekatan hybrid mencoba untuk memanfaatkan kekuatan yang saling melengkapi dari pendekatan *Wrapper* dan *Filter*.

D. Confusion Matrix

Confusion Matrix merupakan sebuah metode untuk evaluasi yang menggunakan tabel matrix seperti pada tabel 1. Pada tabel 1 dapat dilihat bahwa jika dataset terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang

lainnya negatif (Bramer, 2007) [10]. Nilai *accuracy* merupakan persentase jumlah record data yang diklasifikasikan secara benar oleh sebuah algoritma dapat membuat klasifikasi setelah dilakukan pengujian pada hasil klasifikasi tersebut (Han & Kamber, 2006) [11]. Nilai *precision* atau dikenal juga dengan nama *confidence* merupakan proporsi jumlah kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. Sedangkan nilai dari *recall* atau *sensitivity* merupakan proporsi jumlah kasus positif yang sebenarnya yang diprediksi positif secara benar (Powers, 2011) [12].

TABEL I
MODEL CONFUSION MATRIX

Correct classification	Classified as	
	+	-
+	True positive	False negative
-	False positive	True negative

III. METODE PENELITIAN

A. Dataset Penelitian

Dataset yang diolah adalah dataset penundaan penerbangan Bandara A. Yani Semarang

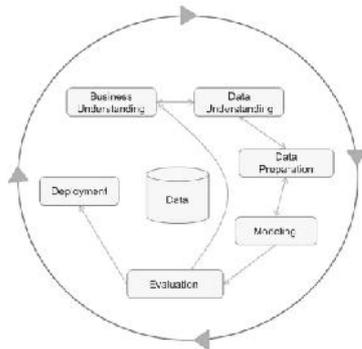
B. Metode

Pada Paper ini akan digunakan metode *Correlation Matrix* untuk mengetahui hubungan antar faktor dalam Penundaan Penerbangan. Kemudian akan dibandingkan 5 algoritma yang telah dikenal secara umum, antara lain *Decision Tree (C4.5)*, *Naive Bayes (NB)*, *K-Nearest Neighbor (KNN)*, *Random Forest (RF)*, dan *Logistic Regression (LogR)*, yang kemudian akan didapatkan algoritma terbaik untuk penentuan penundaan penerbangan.

Selanjutnya digunakan perbandingan 3 metode *Feature Selection* antara lain *Filter (Information Gain)*, *Forward Selection*, dan *Backward Elimination* yang selanjutnya didapatkan metode *Feature Selection* yang terbaik untuk meningkatkan performa dari algoritma yang diusulkan tersebut.

Penelitian dilakukan dengan metode CRISP-DM yang terdiri dari 6 fase:

- Busines Understanding*
- Data Understanding*
- Data Preparation*
- Modeling*
- Evaluation*
- Deployment*



Gbr 1. Alur metode CRISP-DM

- 1) *Business Understanding* merupakan fase awal dari CRISP-DM untuk mengetahui masalah yang akan diselesaikan untuk mencapai tujuan yang diinginkan.
- 2) *Data Understanding* untuk mengenali data yang akan diolah dengan mengumpulkan, mendeskripsikan dan mengevaluasi kualitas data.
- 3) *Data Preparation* diperlukan untuk mempersiapkan data yang akan diolah agar dapat dimodelkan dengan memeriksa apakah data tersebut normal, lengkap dan konsisten sehingga dapat dimodelkan sesuai dengan metode data mining yang akan digunakan.
- 4) *Modeling* merupakan fase untuk memodelkan metode yang digunakan untuk mengolah dataset dalam hal ini menggunakan beberapa metode data mining yaitu Korelasi, Klasifikasi, dan Feature Selection. Pada Korelasi akan dicari hubungan antar faktor sehingga dapat diketahui faktor yang paling berpengaruh dalam menentukan penundaan penerbangan, pada klasifikasi dibandingkan 5 algoritma untuk dipilih yang terbaik yaitu Decision Tree, Naïve Bayes, K-NN, Random Forest, dan Logistic Regression. Sedangkan pada Feature Selection dibandingkan 3 algoritma yaitu Filter Selection, Forward Selection (FS), dan Backward Elimination (BE).
- 5) *Evaluation* digunakan untuk mengevaluasi hasil pengolahan data sehingga dapat diketahui pola / pengetahuan dari hasil pemodelan tersebut.
- 6) *Deployment* (What Next) merupakan penerapan / tindak lanjut yang diperlukan setelah diketahui pengetahuan / pola untuk menyelesaikan masalah untuk mencapai tujuan yang diinginkan.

IV. HASIL DAN PEMBAHASAN

Eksperimen dilakukan pada laptop berbasis Core i5 2.40 GHz CPU, 4 GB RAM dan sistem operasi Windows 10 Enterprise 64-bit. Aplikasi yang digunakan adalah RapidMiner 7.4. Dan berikut ini adalah hasil eksperimen dari tiap fase CRISP-DM.

A. *Business Understanding*

1) Motivasi:

- Bandara Ahmad Yani merupakan bandar udara nasional di Semarang

- Bandara perlu untuk mengetahui faktor-faktor yang berpengaruh dalam menentukan penundaan penerbangan domestik.
 - Bandara memahami bahwa ada beberapa faktor yang mempengaruhi dalam menentukan penundaan penerbangan dan percaya bahwa dengan mengetahui hubungan antar faktor tersebut dapat mengambil keputusan yang lebih baik dalam menentukan penundaan penerbangan.
 - Bandara memilih korelasi sebagai cara untuk model hubungan antar faktor untuk diketahui. Korelasi adalah pengukuran statistik bagaimana kekuatan hubungan antar atribut dalam Dataset.
 - Bandara juga perlu untuk mencari metode klasifikasi yang tepat dalam menentukan penundaan penerbangan.
 - Metode klasifikasi yang tepat dapat membantu bandara dalam menentukan penundaan penerbangan dengan baik.
- 2) **Objektif:**
 - Untuk mencari hubungan antar faktor yang mempengaruhi penundaan penerbangan
 - Untuk mencari metode klasifikasi yang terbaik dengan melakukan perbandingan 5 algoritma klasifikasi penundaan penerbangan
 - Untuk meningkatkan performa dari metode klasifikasi dapat dilakukan dengan menggunakan *feature selection*.

B. *Data Understanding*

Untuk meneliti kebutuhan tersebut, bandara telah mencatat dengan membuat suatu correlation matrix, algoritma klasifikasi dan metode feature selection dari enam atribut

Dengan menggunakan sumber data penundaan penerbangan dari database bandara, dibuat Dataset dengan atribut sebagai berikut:

1. Arah Angin: Arah dari mana angin tersebut bertiup dan dinyatakan dengan sudut kompas (⁰)
2. Suhu: suhu rata-rata di bandara diukur dalam ⁰C
3. Kecepatan Angin: Kecepatan aliran udara (angin) di bandara (knot)
4. Tekanan Udara: Tekanan udara di bandara saat akan penerbangan (mb)
5. Visibility: Jarak pandang yang terjangkau oleh penglihatan (m)
6. Cuaca: Kondisi cuaca di bandara ketika akan dilakukan penerbangan. Cerah (SN), Hujan (RA), Halilintar (TS), Halilintar sekitar bandara (VCTS)

C. *Data Preparation*

Atribut tersebut disesuaikan dengan metode data mining yang akan digunakan yaitu untuk korelasi atributnya bisa numerik atau nominal dan tidak ada label, untuk klasifikasi maka atributnya bisa numerik atau nominal dan labelnya harus dalam bentuk nominal dalam hal ini keterangannya adalah delay atau tidak.

Kemudian setelah dievaluasi kualitas data ternyata datasetnya normal/bersih tidak ada data yang kosong dan sudah konsisten sehingga dapat dimodelkan.

TABEL II

DATASET PENUNDAAN PENERBANGAN BANDARA A. YANI SEMARANG

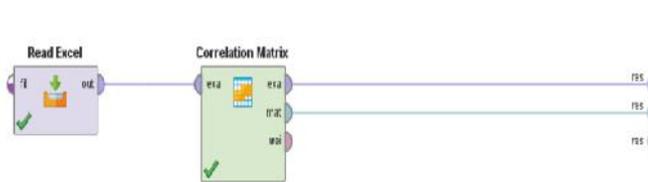
Arah angin	kecepatan angin	Suhu	Tekanan udara	visibility	cuaca	Keterangan
350	10	27,6	1008,4	5000	RA VCTS	TIDAK
330	18	27,1	1009,4	5000	TS RA	TIDAK
330	13	26,9	1008,3	3000	TS RA VCTS	TIDAK
320	16	27,5	1006,7	3000	TS RA	DELAY
330	17	27,7	1006,8	5000	TS RA VCTS	TIDAK
270	14	27,7	1007,5	6000	TS RA	TIDAK
350	11	26,8	1007,3	2000	TS RA	DELAY
310	14	27,3	1006,9	3000	TS RA	TIDAK
300	10	27,7	1006,2	6000	RA	TIDAK
310	18	28,4	1005,7	2600	RA VCTS	DELAY
300	22	28,4	1007,5	4000	RA	TIDAK
300	23	28,6	1008,7	3000	TS RA	TIDAK
330	15	26,3	1008,7	3000	TS RA	TIDAK
300	17	26,9	1009,8	1000	TS RA	DELAY
320	14	25,4	1010,8	1000	RA	DELAY
300	16	26,1	1010,3	4000	RA	TIDAK
300	11	28,1	1009,6	6000	RA	TIDAK
320	18	27,7	1010,2	5000	RA VCTS	TIDAK
310	17	27,2	1010	3000	TS RA VCTS	DELAY
320	15	26,8	1010,2	5000	SN	TIDAK
290	20	27,6	1011	6000	TS RA VCTS	TIDAK
310	23	27,5	1010,6	5000	TS RA	TIDAK
290	19	27,3	1009,9	5000	RA	TIDAK

Apabila diperlihatkan dalam statistik akan tampak seperti gambar 2.

Gbr. 2 Statistik Dataset Penundaan Penerbangan Bersih (clean)

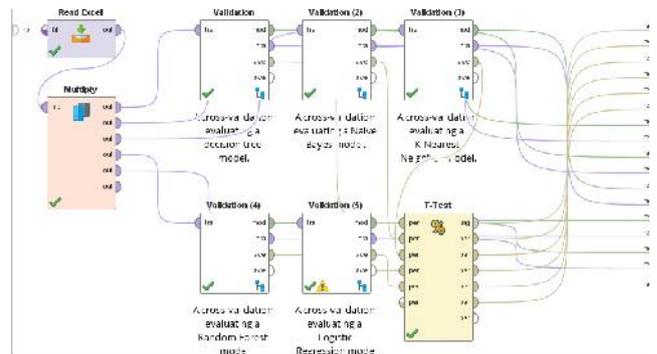
D. Modeling

Merupakan fase pemilihan teknik data mining dengan menentukan algoritma yang akan digunakan. Tool yang digunakan adalah RapidMiner versi 7.4. Untuk mengetahui hubungan antar faktor atribut digunakan Correlation Matrix yang dapat mendeskripsikan bentuk dan kekuatan hubungan antar faktor tersebut.



Gbr3. Model Correlation Matrix

Untuk metode Klasifikasi, dengan membandingkan 5 Algoritma yaitu: *Decision Tree*, *Naive Bayes*, *K-NN*, *Random Forest*, dan *Logistic Regression*, untuk diketahui yang terbaik.



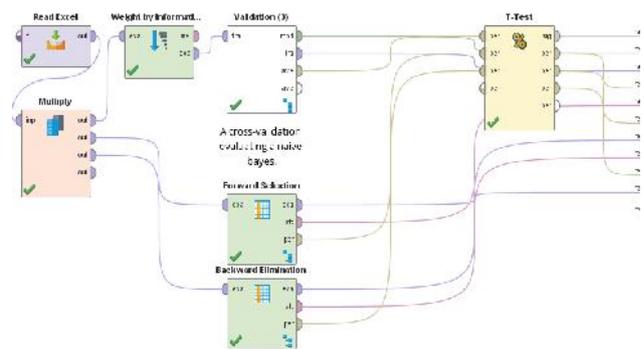
Gbr 4. Model Klasifikasi dengan 5 Algoritma

Digunakan Uji beda (T-Test) Untuk membandingkan kinerja (performa) dari 5 algoritma tersebut. Dengan uji beda (T-test) dapat diketahui Akurasi untuk Klasifikasi dan perbedaan signifikan dari kelima algoritma dari masing-masing metode Klasifikasi tersebut untuk dianalisa sehingga dapat diketahui yang terbaik.

Untuk memperbaiki kinerja (performa) dari masing-masing metode dapat digunakan metode *Feature Selection* yang terdiri dari beberapa jenis, sedangkan yang digunakan dalam penelitian ini adalah :

- Filter (information gain)
- Wrapper (Forward Selection dan Backward Elimination)

Feature Selection dapat memperbaiki performa karena dapat meningkatkan Akurasi. Untuk membandingkan metode *Feature Selection* tersebut juga digunakan uji beda (T-Test). Hasil dari uji beda tersebut digunakan untuk mengetahui metode *feature Selection* yang terbaik.

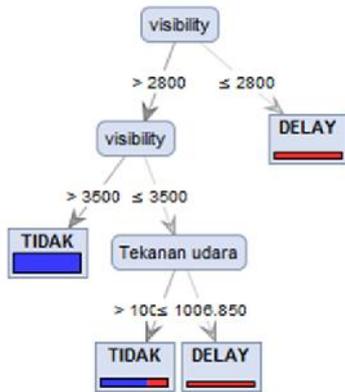


Gbr 5. Model Feature Selection untuk Naive Bayes

E. Evaluasi

Dari hasil pemodelan tersebut diketahui bahwa yang paling berpengaruh dalam menentukan penundaan

penerbangan adalah *visibility* (jarak pandang). Penundaan penerbangan dilakukan apabila jarak pandangnya kurang dari 2800 m. Jarak pandang ini diakibatkan oleh kondisi cuaca yang juga berkaitan dengan suhu, tekanan udara, dan kecepatan angin.

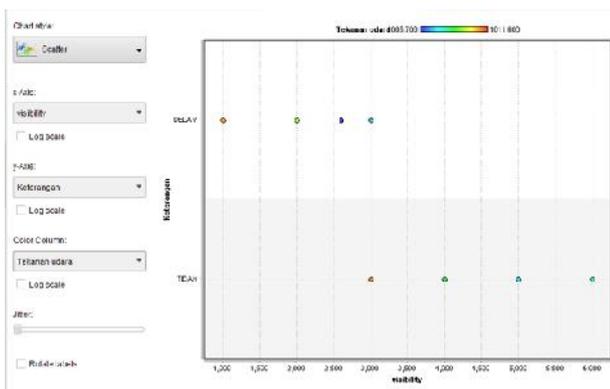


Gbr 6. pola klasifikasi penundaan penerbangan

Pola/pengetahuan yang didapatkan adalah sebagai berikut

- Apabila *visibility* ≤ 2800 maka ditunda, bila tidak kemudian
- Jika *visibility* > 3500 maka diperbolehkan, bila tidak maka kemudian
- Jika tekanan udara ≤ 1006.850 maka ditunda, bila tidak maka diperbolehkan.

Kemudian dari Hasil *Correlation Matrix* dapat diketahui bahwa faktor yang paling berpengaruh dalam menentukan penundaan penerbangan adalah *visibility* (jarak pandang). Sebagaimana telah dijelaskan sebelumnya bahwa semua rencana penerbangan dengan kondisi jarak pandang (*visibility*) yang rendah kurang dari 2800 m ditunda penerbangannya.



Gbr 7. Grafik hubungan antar faktor *visibility*, tekanan udara & keterangan

Kemudian dapat juga diketahui hubungan antar faktor diantaranya adalah hubungan positif (berbanding lurus) seperti pada hubungan antara cuaca dengan tekanan udara, hal ini

dapat diketahui dari nilai *correlation*-nya jika positif maka hubungannya positif (berbanding lurus) jika nilai *correlation*-nya negatif maka hubungannya negatif (berbanding terbalik).



Gbr 8. Correlation (hubungan) positif

Lalu untuk hubungan negatif (berbanding terbalik) yaitu antara keterangan dan *visibility* dimana semakin rendah *visibility*-nya maka penerbangan akan semakin pasti untuk ditunda.

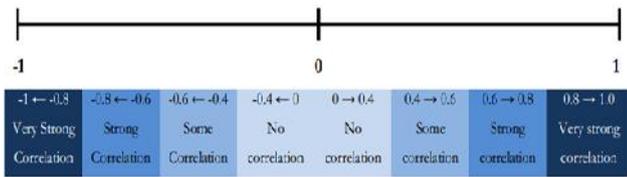


Gbr 9. Correlation (hubungan) negatif

Selain itu dapat diketahui juga kekuatan hubungannya semakin besar nilai *correlation*-nya maka semakin kuat/banyak hubungannya begitu juga sebaliknya semakin kecil nilai *correlation*-nya maka semakin lemah/sedikit hubungannya contohnya adalah hubungan antara keterangan (penundaan) dengan *visibility* memiliki hubungan yang sangat kuat karena penundaan penerbangan sangat tergantung pada *visibility* sehingga nilai *correlation*-nya besar. Sedangkan untuk yang hubungannya lemah/sedikit contohnya adalah hubungan antara cuaca dengan tekanan udara karena memiliki nilai *correlation* yang kecil. Sedangkan yang nilai *correlation*-nya sangat kecil atau bisa dikatakan tidak berhubungan karena nilai *correlation*-nya kurang dari 0.4 seperti yang ditunjukkan pada gambar hasil *correlation matrix* berikut ini

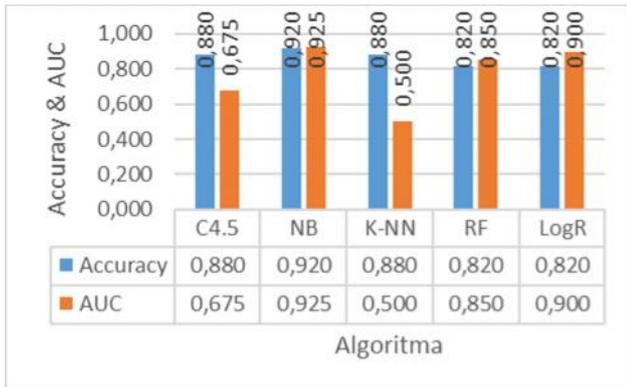
Attributes	Arah angin	kecepatan angin	Suhu	Tekanan udara	<i>visibility</i>	cuaca	Keterangan
Arah angin	1	-0.307	0.043	0.008	-0.018	0.075	-0.073
kecepatan angin	-0.307	1	0.229	0.034	-0.168	-0.227	0.031
Suhu	0.043	0.229	1	-0.231	0.380	0.149	-0.250
Tekanan udara	0.008	0.034	-0.231	1	0.134	0.441	-0.251
<i>visibility</i>	-0.018	-0.168	0.380	0.134	1	0.348	-0.759
cuaca	0.075	-0.227	0.149	0.441	0.348	1	-0.355
Keterangan	-0.073	0.031	-0.250	-0.251	-0.759	-0.355	1

Gbr 10. Tabel hasil *Correlation matrix* hubungan antar faktor atribut



Gbr 11. Kekuatan hubungan antar faktor

Kemudian dari hasil uji beda (T-Test) diketahui bahwa pada Klasifikasi model yang terbaik adalah Naive Bayes (NB) karena memiliki tingkat Akurasi yang tertinggi dan tidak memiliki perbedaan signifikan.



Gbr 12. Accuracy dan AUC pada 5 Algoritma

A	B	C	D	E	F
	0,580 +/- 0,100	0,920 +/- 0,068	0,880 +/- 0,100	0,820 +/- 0,108	0,820 +/- 0,108
0,880 +/- 0,100		0,500	1,000	0,338	0,338
0,920 +/- 0,095			0,509	0,043	0,043
0,880 +/- 0,100				0,338	0,338
0,820 +/- 0,108					1,000
0,820 +/- 0,108					

Gbr 14. Hasil Uji beda (T-Test) Klasifikasi

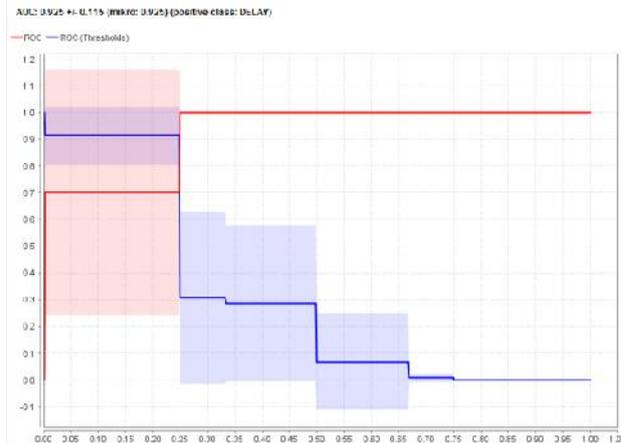
Dari hasil tersebut dapat diketahui bahwa urutan model terbaiknya adalah: 1). NB; 2). C4.5; 3).K-NN; 4). LR; 5.) RF. Untuk mengetahui akurasi dari Naive Bayes ini dapat dilihat dengan *Confusion Matrix* berikut ini :

Table View Plot View

accuracy: 92.00% +/- 0.80% (mikro: 92.00%)

	true TIDAK	true DELAY	class precision
pred. TIDAK	36	1	97.30%
pred. DELAY	3	10	75.92%
class recall	92.31%	90.91%	

Gbr 13. Confusion Matrix NB



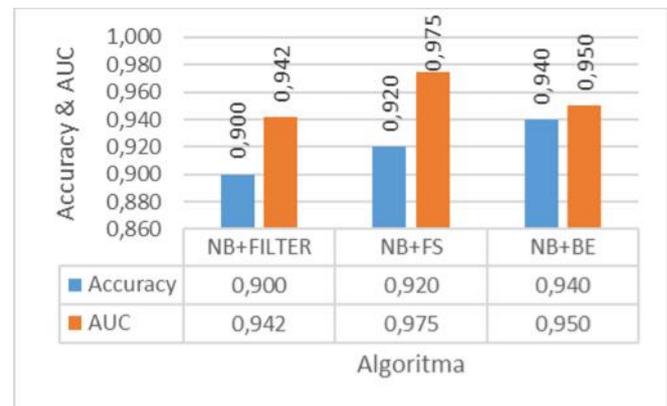
Gbr 14. Kurva ROC-AUC untuk Naive Bayes

Kategori Klasifikasi AUC:

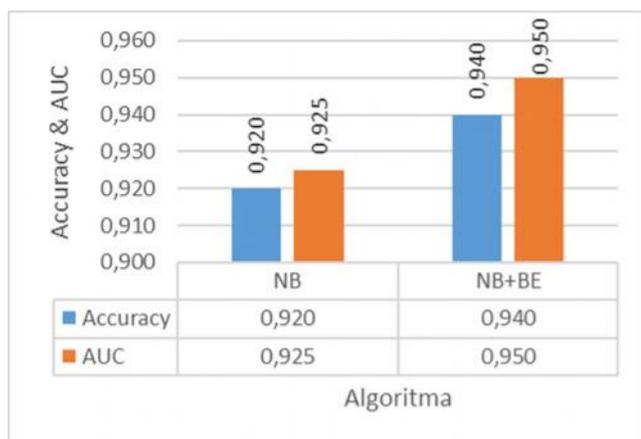
- 0.90 - 1.00 = excellent classification
- 0.80 - 0.90 = good classification
- 0.70 - 0.80 = fair classification
- 0.60 - 0.70 = poor classification
- 0.50 - 0.60 = failure

Dari kurva ROC-AUC model NB memiliki AUC sebesar 0.925 ini berarti termasuk dalam kategori klasifikasi *excellent*.

Hasil metode *Feature Selection* pada klasifikasi yang terbaik adalah Backward Elimination (BE) sehingga model terbaik yang digunakan adalah NB+BE. Dengan Feature Selection ini performanya lebih baik dari sebelumnya karena akurasi meningkat dari 92.00% menjadi 94.00% dan AUC meningkat dari 0.925 menjadi 0.950.



Gbr 15. Accuracy dan AUC Feature Selection pada Algoritma Naive Bayes



Gbr 16. Grafik Accuracy dan AUC pada NB sebelum dan sesudah menggunakan BE

A	B	C	D
	0.900 +/- 0.100	0.920 +/- 0.133	0.940 +/- 0.092
0.900 +/- 0.100		0.708	0.363
0.920 +/- 0.133			0.689
0.940 +/- 0.092			

Gbr 17. Hasil Uji beda (T-Test) Feature Selection untuk NB

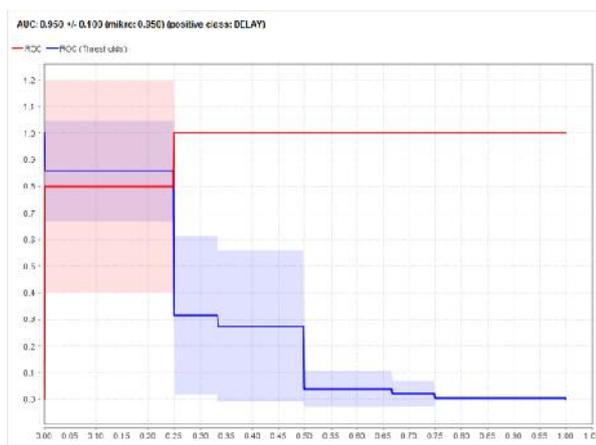
Dari hasil tersebut dapat diketahui bahwa urutan model terbaiknya adalah: 1. NB+BE 2. NB+FS 3. NB+FILTER.

Table View Plot View

accuracy: 94.00% +/- 9.17% (mikroc: 94.00%)

	true TIDAK	true DELAY	class precision
pred. TIDAK	36	0	100.00%
pred. DELAY	3	11	78.57%
class recall	92.31%	100.00%	

Gbr 18. Confusion Matrix NB+BE



Gbr 19. Kurva ROC-AUC untuk NB+BE

Dari kurva ROC-AUC model NB+BE memiliki AUC sebesar 0.950 ini berarti termasuk dalam kategori klasifikasi excellent

F. Deployment

Telah dihasilkan suatu informasi, dan pola pengetahuan baru dalam proses data mining. Pola pengetahuan tersebut didapat dari metode Korelasi, Klasifikasi, dan *Feature Selection* untuk menentukan penundaan penerbangan berdasarkan dataset penundaan penerbangan pada Bandara A. Yani Semarang. Untuk atribut lain yang tidak terlalu berpengaruh dalam klasifikasi tersebut dapat dihilangkan seperti arah dan kecepatan angin. Atribut tersebut dapat diganti dengan yang lebih berpengaruh misalnya ketebalan kabut atau awan, dan intensitas cahaya.

Klasifikasi penundaan penerbangan sangat tergantung dari jarak pandang akibat kondisi cuaca. Keakuratan Klasifikasi dapat ditingkatkan dengan menggunakan *Backward Elimination* sehingga dapat menghasilkan keputusan klasifikasi yang lebih akurat. Pengetahuan yang diperoleh dapat digunakan sebagai dasar dalam mengambil keputusan untuk menentukan penundaan penerbangan sehingga dapat mengurangi terjadinya kecelakaan pesawat.

V. KESIMPULAN

Penelitian dapat menggunakan beberapa peran data mining yaitu : Korelasi, Klasifikasi, dan *Feature Selection*

1. Hubungan Antar Faktor Penundaan Penerbangan
 - Faktor yang paling mempengaruhi penundaan penerbangan adalah *visibility* (jarak pandang).
 - Faktor lain tingkat korelasinya sangat kecil terhadap penundaan penerbangan.
 - Apabila jarak pandang kurang dari 2800 m maka penerbangan tidak dapat dilakukan atau ditunda.
 - Apabila jarak pandangnya antara 2800 m dan 3500 m maka dilihat tekanan udaranya, apabila tekanan udaranya ≤ 1006.850 maka ditunda, bila tidak maka diperbolehkan.
 - Akan tetapi walaupun cuaca buruk kalau *visibility* (jarak pandangnya) lebih dari 3500 m maka penerbangan masih memungkinkan untuk dilakukan.
2. Perbandingan 5 Algoritma Klasifikasi Penundaan Penerbangan
 - Algoritma terbaiknya adalah *Naive Bayes* dengan Akurasi 92.00 %, AUC 0.925, dan tidak memiliki perbedaan signifikan.
 - Sehingga algoritma *Naive Bayes* dapat digunakan untuk klasifikasi dalam menentukan penundaan penerbangan dengan baik.
3. Peningkatan performa dengan *Feature Selection*
 - *Feature Selection* dapat mengurangi faktor/atribut yang tidak terlalu berpengaruh sehingga dapat meningkatkan Akurasi dan AUC.

- Sedangkan metode *Feature Selection* yang terbaik untuk NB pada penelitian ini adalah *Backward Elimination* (BE) modelnya menjadi NB+BE.
- Metode dengan model NB+BE memiliki tingkat akurasi yang tinggi sehingga hasilnya cukup akurat dengan demikian metode ini dapat digunakan sebagai rekomendasi dalam membantu mengambil keputusan yang tepat untuk menentukan penundaan penerbangan.

Untuk penelitian mendatang dapat menggunakan atribut yang berpengaruh lainnya misalnya ketebalan kabut atau awan, dan intensitas cahaya. Untuk memperoleh hasil penentuan penundaan penerbangan yang lebih baik dapat dilakukan penelitian dengan menggunakan algoritma lain yang lebih baru dan lebih baik. Selanjutnya untuk lebih meningkatkan performa dari Algoritma tersebut dapat dilakukan dengan menggunakan metode *Feature Selection* atau *Feature Extraction* lain yang dapat meningkatkan performa algoritma menjadi lebih cepat dan lebih akurat.

DAFTAR PUSTAKA

- [1] Soejadi Wh, "Manfaat Dan Bahaya Cuaca Dalam Penerbangan,," <http://pustakacuaca.blogspot.com/2010/08/manfaat-dan-bahaya-cuaca-dalam.htm>. [Accessed 30 Januari 2014].
- [2] Kusriani,&Luthfi, E. T. (2009). *Algoritma Data Mining*. Yogyakarta: Andi Publishing.
- [3] Puspita, E.S. and Yulianti, L., 2016. Perancangan Sistem Peramalan Cuaca Berbasis Logika Fuzzy. *Media Infotama*, 12(1).
- [4] Endah Sari, N. and Sukirman, E., 2012. Prediksi Cuaca Berbasis Logika Fuzzy untuk Rekomendasi Penerbangan di Bandar Udara Raja Haji Fisabilillah.
- [5] Yuniar, R.J., Rahadi, D. and Setyawati, O., 2013. Perbaikan Metode Prakiraan Cuaca Bandara Abdulrahman Saleh dengan Algoritma Neural Network Backpropagation. *Jurnal EECCIS*, 7(1), pp.65-70.
- [6] Abidin, A.Z.Z., 2015, July. Implementasi Algoritma C 4.5 untuk Menentukan Tingkat Bahaya Tsunami. In *Seminar Nasional Informatika (SEMNASIF)* (Vol. 1, No. 1).
- [7] Tasikmalaya, A.B. and Wati, R., 2016. Penerapan Algoritma Genetika Untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan Menggunakan Naive Bayes. *Evolusi*, 4(1).
- [8] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques Third Edition*, Elsevier, 2012.
- [9] Daniel T. Larose, *Discovering Knowledge in Data: an Introduction to Data Mining*, John Wiley & Sons, 2005.
- [10] Bramer, Max. 2007. *Principles of Data Mining*. London: Springer. ISBN-10: 1-84628-765-0, ISBN-13: 978-1-84628-765-7.
- [11] Han, J. and Kamber, M., 2006. *Data Mining: Concepts and Techniques*, University of Illinois at Urbana-Champaign.
- [12] Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.