JuSIN
Jurnal Sistem Informasi
Website: ojs.itb-ad.ac.id/index.php/JUSIN
E-ISSN: 2797-8516

IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI KETERLAMBATAN JAM MASUK KERJA MENGGUNAKAN ALGORITMA KLASIFIKASI

Saeful Bahri[1] [(*)]

[1]*ITB Ahmad Dahlan, Jakarta*

Abstract

The development of computer technology now is really fast. The computer is not only used as a tool to complete the work of humans but can also run applications designed to access information quickly. Application of expert system application is moved to a computer expert knowledge. So the computer could also resolve the problem as usual is done by specialists. Expert system has been widely developed in various fields, including the field of animal husbandry. Chicken breeders usually get the chicken disease information from the livestock extension officers. But sometimes often constrained because the number of extension officers of the cattle that are not evenly distributed in each area. Therefore, the breeder can access information about diseases of chicken with the help of the method of Forward Chaining and Certainty Factor into the making of the application. The resulting application is an expert system for detecting diseases of chickens that could be used to assist farmers in obtaining information about the disease of chickens and handling solutions.

Kata Kunci: *Data mining, delays, classification algorithms*

Juli – Desember 2020, Vol 1 (1) : hlm 11-20

[(*)] Korespondensi: *mr.saeful.bahri@gmail.com (Saeful Bahri)*

PENDAHULUAN

Karyawan menjadi salah satu faktor penting yang dapat menentukan keberhasilan tujuan perusahaan. Hal ini terjadi karena karyawan adalah penggerak utama roda kegiatan sebuah perusahaan. Dengan kata lain, karyawan menjadi penguat fungsi-fungsi organisasi dalam sebuah perusahaan.

Dalam melakukan kegiatan, karyawan memerlukan petunjuk kerja atau pemberitahuan bagaimana melaksanakan sebuah pekerjaan dari perusahaan agar pelaksanaan pekerjaan tersebut sesuai dengan perencanaan yang telah dibuat perusahaan, hal tersebut harus didukung dengan aturan yang dibuat perusahaan agar tercipta disiplin kerja dan tidak terjadi penyimpangan.

Disiplin kerja akan tercipta dengan baik bila perusahaan mampu menciptakan dan mengimplementasikan aturan serta konsekuensi kerja terhadap karyawannya. Hal tersebut berarti bahwa perusahaan harus mensosialisasikan aturan-aturan yang telah dibuat kepada karyawan sampai karyawan memahami dan dapat melaksanakan aturan-aturan tersebut.

Hal ini dilakukan untuk mencegah terjadinya penyimpangan kerja sehingga permasalahan disiplin kerja dapat diatasi. Bentuk dari permasalahan ini dapat kita lihat dalam berbagai bentuk seperti sering datang terlambat pada jam masuk kerja.

Datang terlambat kerja adalah datang lewat dari waktu yang sudah di tentukan oleh perusahaan. Keterlambatan masuk kerja karyawan akan menghambat proses bisnis perusahaan menjadi tertunda, dari sisi karyawan, keterlambatan akan membawa efek nagatif bagi karyawan tersebut, hal ini akan mempengaruhi key performance indicator yang buruk, jenjang karir yang tidak bagus, sampai dengan pemotongan gaji.

Dari permasalahan tersebut, diperlukan sebuah model yang bisa membantu perusahaan untuk memperbaiki kinerja jam kerja karyawan, untuk mengetahui faktor-faktor yang paling berpengaruh terhadap keterlambatan jam masuk kerja, dengan menemukan pola dari setiap keterlambatan jam masuk kerja, pada penelitian ini dilakukan modeling menggunakan tiga algoritma klasifikasi *decision tree, naïve bayes*, dan *k-nearest neighbor* dengan tehnik data mining.

Data mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstrasi pengetahuan (*knowledge*) secara otomatis. Definisi lainnya adalah pembelajaran berbasis induksi (*induction-based learning*) proses pembentukan definisi-definisi konsep umum yang dilakukan dengan cara mengobservasi contoh-contoh spesifik (Gata & Purnomo, 2017).

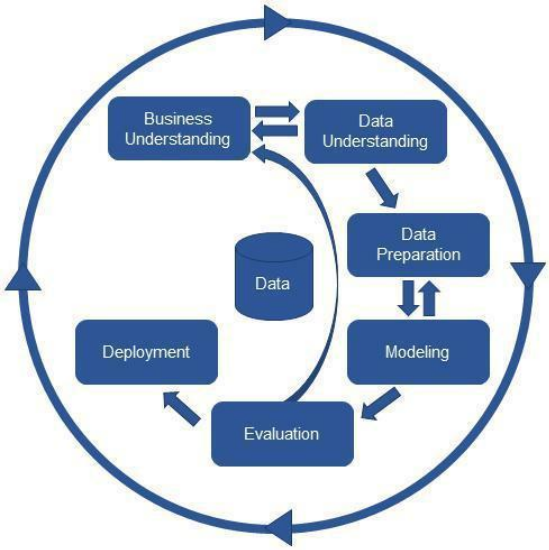
Penelitian menggunakan tehnik data mining ini mengacu kepada penelitian-peneltian sebelumnya seperti: Komparasi model klasifikasi algoritma keterlambatan siswa masuk sekolah (Amirulloh, 2017), pada penelitian ini faktor yang paling mempengaruhi keterlambatan siswa masuk sekolah adalah variabel jam berangkat, dan algoritma K-NN menghasilkan performa yang paling baik diantara algoritma lainnya dengan akurasi 87,38%.

Pada penelitian lain (Muqorobin et al., 2019), dengan judul Optimasi metode *naïve bayes* dengan *feature selection information gain* untuk prediksi Keterlambatan pembayaran sumbangan pembinaan pendidikan sekolah, pada penelitian ini menghasilkan metode algoritma *Information Gain* dengan *Naïve Bayes* untuk prediksi keterlambatan pembayaran SPP Sekolah diperoleh hasil akurasi paling baik dengan akurasi 90%.

METODE

Metode penelitian yang digunakan CRISP-DM (*Cross-Industry Standard Process for Data Mining*) merupakan suatu konsorsium perusahaan yang didirikan oleh Komisi Eropa pada

tahun 1996 dan telah ditetapkan sebagai proses standar dalam data mining yang dapat diaplikasikan di berbagai sektor industri. Gambar menjelaskan tentang siklus hidup pengembangan data mining (Bahri & Indrajit, 1846).



Gambar 2.1 Proses *Data Mining* Menurut *CRISP-DM* (Suherman, 2019).

Tahapan-tahapan data mining (Astuti, 2019):

1. *Business Understanding*

Tahap pertama adalah memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menterjemahkan pengetahuan ini ke dalam pendefinisian masalah pada data mining. Selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut.

2. *Data Understanding*

Tahap ini dimulai dengan pengumpulan data yang kemudian akan dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi adanya bagian yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3. *Data Preparation*

Tahap ini meliputi semua kegiatan untuk membangun dataset akhir (data yang akan diproses pada tahap pemodelan) dari data mentah. Tahap ini dapat diulang beberapa kali. Pada tahap ini juga mencakup pemilihan tabel, *record*, dan atribut-atribut data, termasuk proses pembersihan dan transformasi data untuk kemudian dijadikan masukan dalam tahap pemodelan.

4. *Modelling*

Dalam tahap ini dilakukan pengujian dari variabel yang diguakan untuk mendapatkan nilai yang optimal, dan dilakukan pemilihan model dan penerapan menggunakan teknik data mining. Data mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran Komputer (*machine learning*) untuk menganalisis dan mengekstrasi pengetahuan (*knowledge*) secara otomatis. Definisi lainnya adalah pembelajaran berbasis induksi (*induction-based learning*) proses pembentukan definisi-definisi konsep umum yang dilakukan dengan cara mengobservasi contoh-contoh spesifik (Mustafa et al., 2018).

Kemajuan dalam bidang *data mining* didorong oleh beberapa faktor antara lain (Meilina, 2015):

a. Pertumbuhan yang cepat dalam kumpulan data.
b. Penyimpanan data dalam data *warehouse*, sehingga seluruh perusahaan memiliki akses ke dalam *database* yang andal.

c. Adanya peningkatan akses data melalui navigasi web dan intranet.
d. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
e. Perkembangan teknologi perangkat lunak untuk *data mining* (ketersediaan teknologi).
f. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan

Model yang digunakan pada penelitian ini menggunakan tiga algoritma klasifikasi: *decision tree, naïve bayes,* dan *k-nearest neighbor.*

a. *Decision Tree*

Decision Tree merupakan algortima yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metodeklasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusa mengubah fakta yang sangat besar menjadi pohon keputusanyang merepresentasikan aturan. Aturan dapat dengan mudah dipahamidengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentukbahasa basis data seperti Structured Query Language untuk mencari record pada kategori tertentu (Yu et al., 2007). Cara algoritma C4.5 untuk membangun pohon keputusan yaitu:

- Pilih atribut sebagai akar.
- Buat cabang untuk masing-masing nilai
- Bagi kasus dalam cabang.
- Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada.Untuk menghitung gain digunakan rumus seperti tertera dalam rumus dibawah ini.

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} * Entropy(Si)$$

Dengan:
S : Himpunan kasus
A : Atribut
n : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke i
|S| : Jumlah kasus dalam S

b. *Naïve Bayes*

Naïve bayes merupakan salah satu algoritma dalam teknik data mining yang menerapkan teori Bayes dalam klasifikasi. Teorema keputusan Bayes adalah adalah pendekatan statistik yang fundamental dalam pengenalan pola *(pattern recoginition)*. Naive bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara konditional saling bebas jika diberikan nilai *output*. Dengan kata lain, diberikan nilai *output*, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Dengan memasukkan Persamaan 1 ke Persamaan 2 akan diperoleh pendekatan yang digunakan *dalam* naïve bayes (Ridwan et al., 2013).

c. *K-Nearest Neighbor.*

K-Nearest Neighbor adalah suatu metode yang menggunakan algoritma supervised learning dimana hasil dari instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori k-tetangga terdekat. Algoritma *k-Nearest Neighbor* menggunakan *Neighborhood Classification* sebagai nilai prediksi dari nilai *instance* yang baru.

Untuk mencari dekat atau jauhnya jarak antar titik pada kelas k biasanya dihitung menggunakan jarak Euclidean . Jarak *Euclidean* adalah formula untuk mencari jarak antara 2 titik dalam ruang dua dimensi (Yustanti, 2012).

5. *Evaluation*

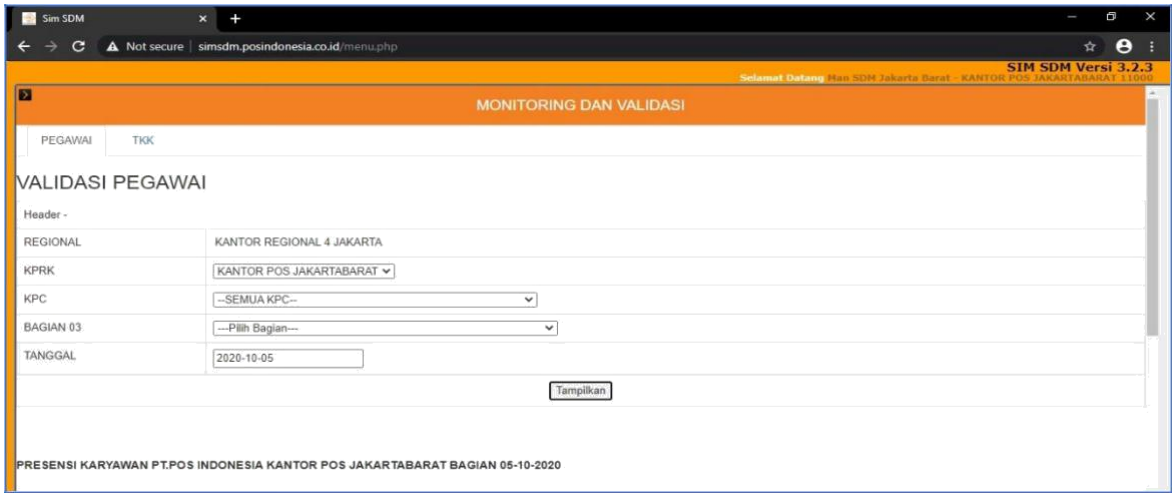
Pada tahap ini dilakukan evaluasi untuk mengetahui akurasi dari model yang diusulkan. Selain itu dilakukan validasi dengan teknik 10 *fold Cross Validation, K-fold Cross Validation* yang merupakan teknik validasi dengan membagi data awal secara acak kedalam *k* bagian yang saling terpisah atau "*fold".* Evaluasi dan validasi menggunakan metode *confusion matrix* dan kurva ROC. Grafik *Receiver Operating Characteristics* (ROC) adalah teknik untuk memvisualisasikan, mengorganisasikan dan memilih pengklasifikasi berdasarkan kinerja setiap algoritma. Kurva ROC digunakan untuk mengukur nilai *Area Under Curve* (AUC). Nilai akurasi algortima diukur menggunakan *confusion matrix* dan hasil peritungan akan ditampilkan dalam bentuk kurva ROC.

6. *Deployment*

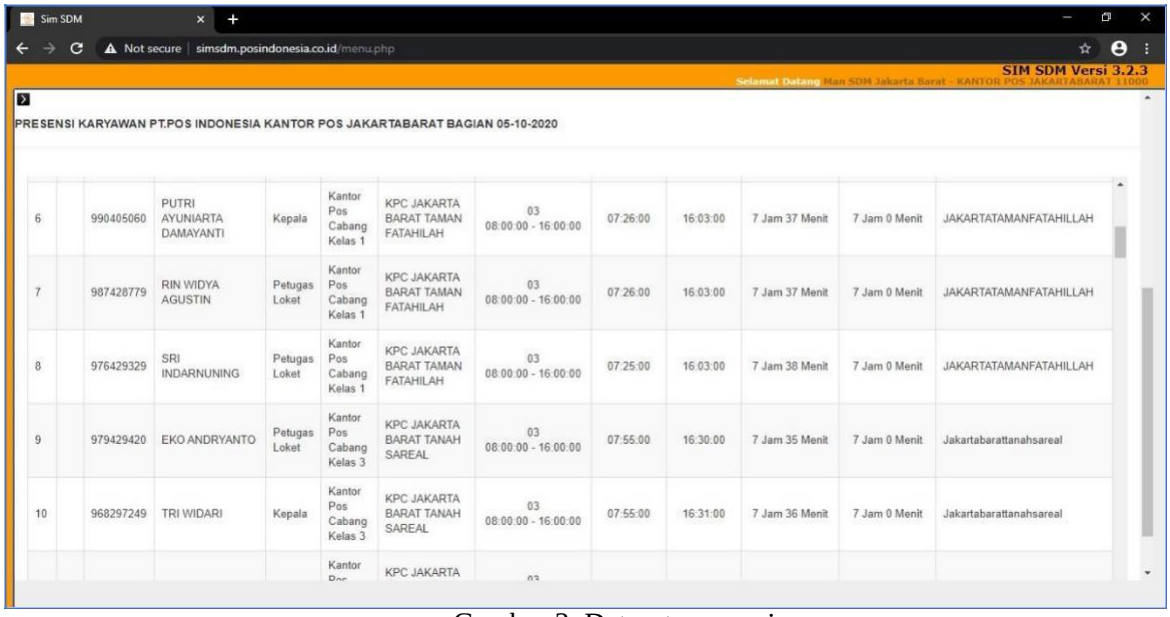
Pada tahap ini, pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Tahap deployment dapat berupa pembuatan laporan sederhana atau mengimplementasikan proses data mining yang berulang dalam perusahaan. Pada banyak kasus, tahap *deployment* melibatkan konsumen, di samping analis data, karena sangat penting bagi konsumen untuk memahami tindakan apa yang harus dilakukan untuk menggunakan model yang telah dibuat.

HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah data presensi karyawan pada PT. Pos Indonesia (Persero) Kantor Pos Jakarta Barat, proses *crawling* data di ambil pada system *https://simsdm.posindonesia.co.id/presensi* dengan jumlah data presensi 164 *record*.



Gambar 2. Crawling dataset pada sistem presensi



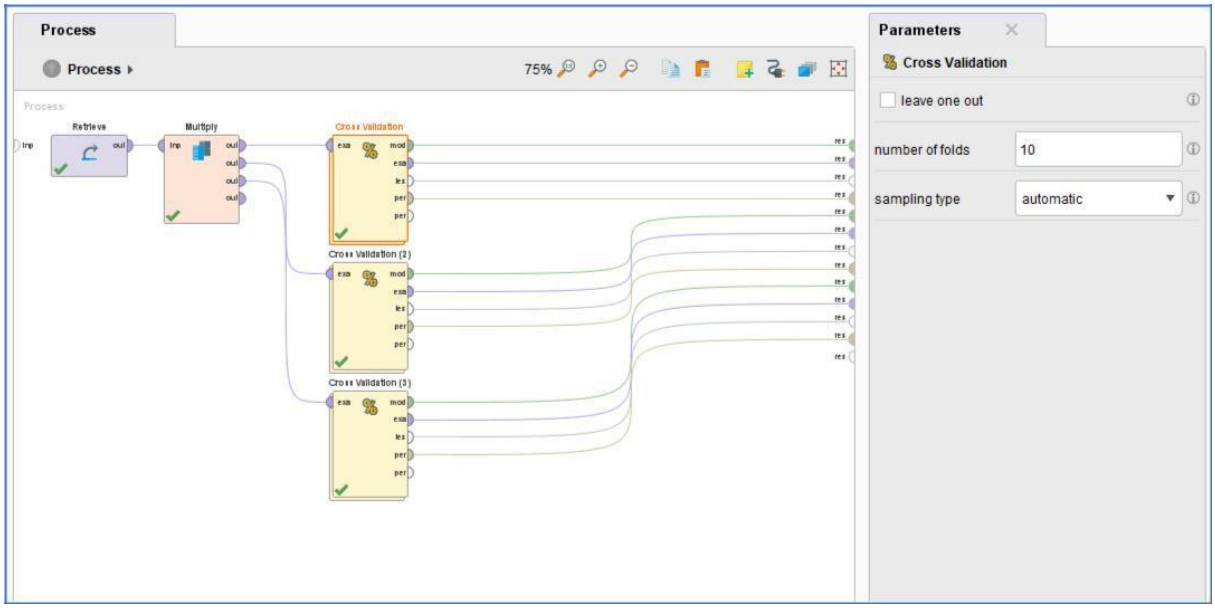
Gambar 3. Dataset presensi

Setelah proses pengumpulan data, data dipersiapkan untuk dilakukan tahap pengujian. Pada tahap ini mencakup pemilihan tabel, record, dan variabel-variabel data, termasuk proses pembersihan variabel yang tidak akan mempengaruhi hasil akhir.

Nama	Jam Masuk	Jam Bangun	Jam Berangkat	Jarak Km2	Transportasi	Tingkat Kemacetan	Cuaca	Label
A1	8:00	6:30	7:30	14.4	Motor	Sedang	Cerah	Terlambat
A2	8:00	5:00	6:15	10.8	Motor	Sedang	Hujan	Tidak Terlambat
A3	8:00	5:00	7:00	9.2	Motor	Lancar	Cerah	Tidak Terlambat
A4	8:00	4:45	6:30	16.5	Motor	Sedang	Hujan	Tidak Terlambat
A5	8:00	4:30	5:00	47.8	Kreta	Lancar	Hujan	Tidak Terlambat
A6	8:00	5:15	7:00	3.1	Motor	Lancar	Cerah	Tidak Terlambat
A7	8:00	6:15	7:30	9.2	Motor	Sedang	Cerah	Terlambat
A8	8:00	5:40	7:15	6.3	Motor	Sedang	Cerah	Tidak Terlambat

Tabel 1. Dataset

1. Modeling



Gambar 4. Model pengujian

Proses pemodelan di atas merupakan proses pengujian model *decision tree, naïve bayes,* dan *k-nearest neighbor* setelah melalui proses *data preparation, dilanjutkan* proses *Set Role* yang berfungsi untuk menentukan *label*, kemudian menggunakan validasi 10 *fold cross Validation* dalam proses *training*, sedangkan untuk proses *testing* menggunakan *apply model* dan *performance.* Proses *training* dan *testing* untuk mendapatkan *confusion matrix* yaitu nilai tingkat *accuracy*, *class precision*, *class recall,* dan kurva ROC yaitu nilai AUC.

2. Hasil pengujian algotirma *decision tree*

accuracy: 73.64% +/- 8.58% (micro average: 73.78%)

	true Terlambat	true Tidak Terlambat	class precision
pred. Terlambat	54	30	64.29%
pred. Tidak Terlambat	13	67	83.75%
class recall	80.60%	69.07%	

Dari hasil pebgujian model decision tree di atas, didapatkan hasil akurasi 73.64% yang artinya tingkat akurasi data sudah baik.

3. Hasil pengujian algotirma *naïve bayes*

accuracy: 40.26% +/- 4.28% (micro average: 40.24%)

	true Terlambat	true Tidak Terlambat	class precision
pred. Terlambat	66	97	40.49%
pred. Tidak Terlambat	1	0	0.00%
class recall	98.51%	0.00%	

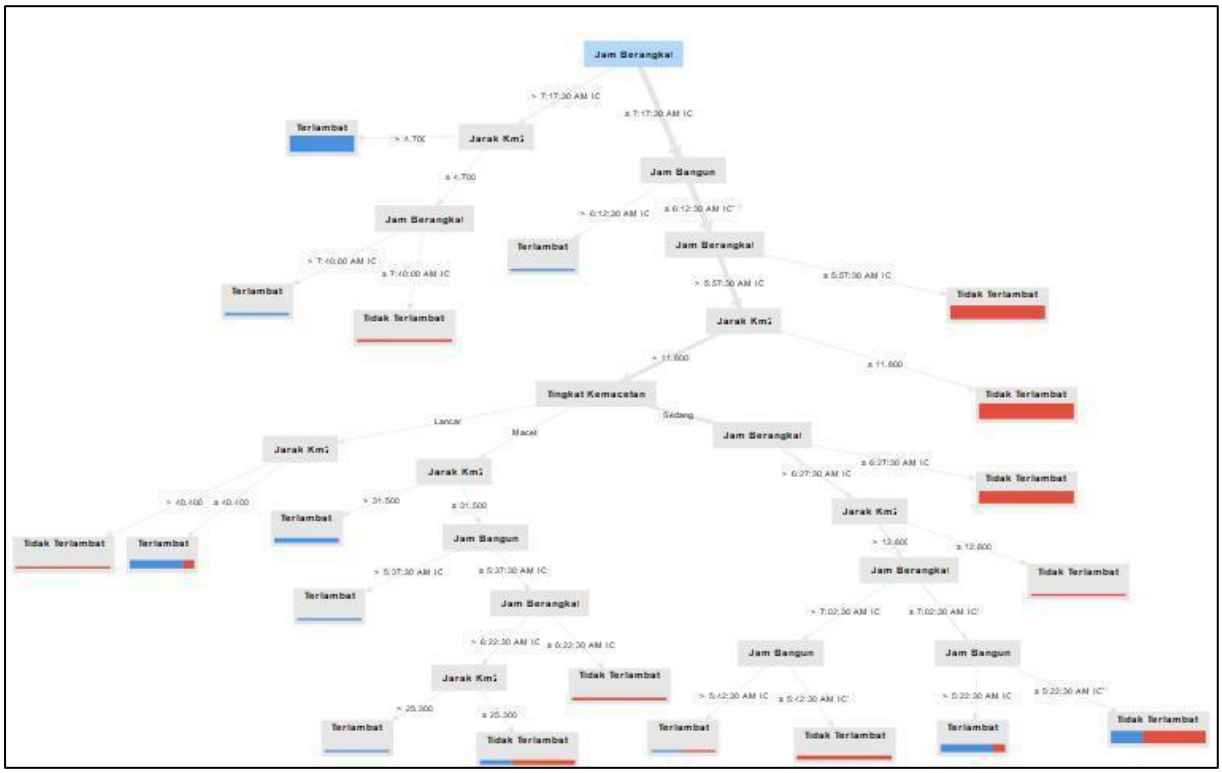
Dari hasil pebgujian model naïve bayes di atas, didapatkan hasil akurasi 40.28% yang artinya tingkat akurasi data buruk.

4. Hasil pengujian algotirma k-*nearest neighbor*

accuracy: 73.09% +/- 6.93% (micro average: 73.17%)

	true Terlambat	true Tidak Terlambat	class precision
pred. Terlambat	42	19	68.85%
pred. Tidak Terlambat	25	78	75.73%
class recall	62.69%	80.41%	

Dari hasil pebgujian model k-*nearest neighbor* di atas, didapatkan hasil akurasi 73.09% yang artinya tingkat akurasi data buruk.



Gambar 5. Pohon keputusan

Output pohon keputusan yang dihasilkan diatas, menunjukan faktor yang paling mempengaruhi keterlambatan jam masuk kerja adalah variabel "jam berangkat".

a. Apabila jam berangkat >07:17:30 akan ditentukan oleh variabel jarak, jarak antara rumah dan tempat bekerja. Apabila jam berangkat <07:17:30 maka ditentukan oleh variabel jam bangun tidur.
b. Variabel jarak, aapabila jarak rumah dengan kantor >4,7 km maka akan terlambat, apabila jarak <4,7 km maka akan di tentukan jam berangkat kerja.
c. Variabel jam bangun tidur, apabila jam bangun >06:12:30 maka akan terlambat masuk kerja, apabila jam bangun <06:12:30 akan di tentukan variabel jam berangkat kerja.

Model	Decision Tree	Naïve Bayes	k-NN
Accuracy	73.64%	40.26%	73.09%
AUC	0.798	0.782	0.759

Tabel 2. Komparasi Akurasi dan AUC

Dari tabel diatas dapat disimpulkan bahwa nilai akurasi dan AUC dari Decission Tree lebih baik dibandingkan yang lainnya. Jika dilihat dari akurasi, perbedaan dari algoritma *decision tree* dan *k-nearest neighbor* tidak terlalu jauh, dan nilai akurasi yang paling kecil adalah Naive Bayes 40.26%.

KESIMPULAN

1. Dari hasil pengujian tiga algoritma yang digunakan, kinerja algoritma *decision tree* menjadi yang terbaik dibandingkan dengan algoritma *k-NN* dan *Naïve bayes* dengan tingkat akurasi 73.64% dan nilai AUC 0.798.
2. Output pohon keputusan yang dihasilkan dari dataset yang dugunakan, menunjukan faktor yang paling mempengaruhi keterlambatan jam masuk kerja adalah variabel "jam berangkat".
3. Dari hasil peneltian ini, diharapkan mampu memberikan solusi bagi perusahaan, untuk meningkatkan kedisiplinan karyawan, dengan melihat faktor-faktor yang mempengaruhi keterlambatan jam masuk kerja.

DAFTAR PUSTAKA

Amirulloh, I. (2017). *Komparasi Model Klasifikasi Algoritma Keterlambatan Siswa Masuk Sekolah*. *November*, 1–2.

Astuti, D. (2019). Penentuan Strategi Promosi Usaha Mikro Kecil Dan Menengah (UMKM) Menggunakan Metode CRISP-DM dengan Algoritma K-Means Clustering. *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, *1*(2), 60–72. https://doi.org/10.20895/inista.v1i2.71

Bahri, S., & Indrajit, R. E. (1846). *Peran Business Intellegence Dalam Peningkatan Penjualan Produk Jasa Keuangan (Pos Payment) Pada Kantor Pos Jakarta Barat*. *November 2017*, 1–2.

Gata, W., & Purnomo. (2017). Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content SMS-Gateway. *Jurnal Format*, *6*(5), 1–5.

Meilina, P. (2015). Penerapan Data Mining Dengan Metode Kalsifikasi MENGGUNAKAN DECISION TREE DAN REGRESI. *Jurnal Teknologi*, *7*(1), 11–20. https://doi.org/10.24853/jurtek.7.1.11-20

Muqorobin, M., Kusrini, K., & Luthfi, E. T. (2019). Optimasi Metode Naive Bayes Dengan Feature Selection Information Gain Untuk Prediksi Keterlambatan Pembayaran Spp Sekolah. *Jurnal Ilmiah SINUS*, *17*(1), 1. https://doi.org/10.30646/sinus.v17i1.378

Mustafa, M. S., Ramadhan, M. R., & Thenata, A. P. (2018). Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Creative Information Technology Journal*, *4*(2), 151. https://doi.org/10.24076/citec.2017v4i2.106

Ridwan, M., Suyono, H., & Sarosa, M. (2013). Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Eeccis*, *7*(1), 59–64. https://doi.org/10.1038/hdy.2009.180

Suherman, E. (2019). *Agglomerative Hierarchical Clustering Dengan Berbagai Pengukuran Jarak Dalam Mengklaster Daerah Berdasarkan Tingkat Kemiskinan*. *5*(1), 978–979.

Yu, L., Chen, G., Koronios, A., Zhu, S., & Guo, X. (2007). Application and Comparison of Classification Techniques in Controlling Credit Risk. *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications*, 2007–2007.

Yustanti, W. (2012). Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah. *Jurnal Matematika Statistika Dan Komputasi*, *9*(1), 57–68.