JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

PERBANDINGAN ALGORITMA KNN, *DECISION TREE*, DAN *RANDOM FOREST* PADA DATA *IMBALANCED CLASS* UNTUK KLASIFIKASI PROMOSI KARYAWAN

LOUIS MADAERDO SOTARJUA[1], DIAN BUDHI SANTOSO[2]

[1,2] Program Studi Teknik Elektro
Fakultas Teknik
Universitas Singaperbangsa Karawang
Jl. HS Ronggo Waluyo, Puseurjaya, Kec. Telukjambe
Timur, Kab. Karawang, Jawa Barat. 41361
E-mail: [1]louis.madaerdo18148@student.unsika.ac.id, [2]dian.budhi@ft.unsika.ac.id

ABSTRAK

Di era *big data*, promosi karyawan dapat dilakukan dengan menggunakan algoritma *machine learning* yang akan mengklasifikasikan dan memprediksi data secara cepat dan konsisten. Pada penelitian ini dilakukan proses klasifikasi dengan algoritma *machine learning*, diantaranya *K-Nearest Neighbor* (KNN), *Decision Tree* dan *Random Forest*. Penelitian bertujuan untuk menganalisa performa model *machine learning* pada data klasifikasi karyawan, data yang digunakan pada penelitian ini merupakan data *imbalanced class* sehingga dilakukan teknik *Synthetic Minority Over-Sampling Technique* (SMOTE). Berdasarkan hasil uji coba, algoritma KNN merupakan algoritma yang memiliki performa terbaik dan tidak mengalami *underfitting*, maupun *overfitting*.

Kata kunci: *Decision Tree*, KNN, *Machine Learning*, *Random Forest*, SMOTE

I. PENDAHULUAN

Sumber Daya Manusia (SDM) merupakan peran utama dalam berjalannya suatu organisasi atau perusahaan, sehingga organisasi atau perusahaan harus memiliki SDM yang berkualitas dalam hal pengetahuan, keterampilan dan kemampuan serta memiliki kemauan untuk mengelola perusahaan secara baik sehingga terciptanya peningkatan kinerja karyawan (Desthiani, 2018). Karwayan dianggap sebagai aset dan investasi bagi suatu perusahaan atau organisasi sehingga akan menghasilkan imbal hasil (*return*) keuntungan bagi perusahaan atau organisasi (Putri, 2013). Peningkatan kinerja karyawan ini juga akan berpengaruh pada kegiatan promosi karyawan, semakin baik promosi karyawan dilakukan, akan

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

semakin baik pula kinerja karyawan dalam perusahaan atau organisasi tersebut (Ningsi, 2015).

Dalam perkembangan teknologi dan era *big data* saat ini, tentunya pekerjaan-pekerjaan yang kompleks akan menjadi mudah dan cepat untuk diselesaikan. Proses penyeleksian dan promosi karyawan pada suatu perusahaan akan dengan mudah dikerjakan jika perusahaan memiliki *database* yang dapat dimanfaatkan untuk diolah ke dalam suatu proses *data mining*. Data mining merupakan suatu proses yang dilakukan untuk mencari pola data maupun informasi dalam *raw data* menggunakan algoritma *machine learning*, sehingga menghasilkan suatu keputusan yang tepat (Mardi, 2017).

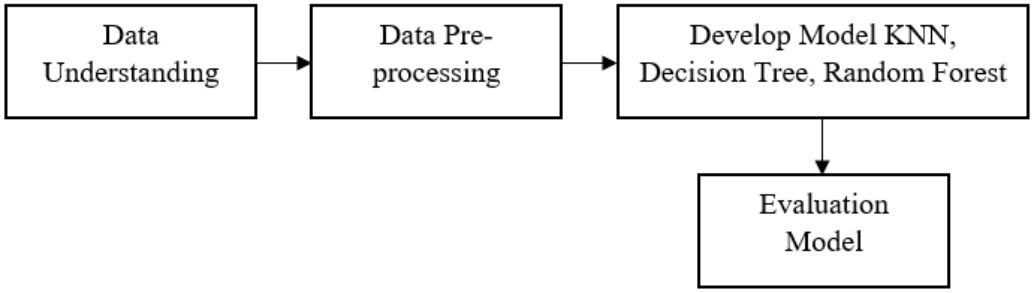
Analisa klasifikasi data adalah proses menemukan model yang menjelaskan pengklasifikasian data yang digunakan untuk memprediksi kelas, label maupun kategori dari sebuah *raw data* (Krisandi, 2013). Terdapat beberapa algoritma untuk klasifikasi dalam *data mining*, diantaranya *K-Nearest Neighbor* (KNN), *Decision Tree*, *Random Forest*, *Support Vector Machines* (SVM), *Recurrent Neural Network* (RNN), dan *Convolutional Neural Network* (CNN). Pada penelitian kali ini digunakan algoritma KNN, *Decision Tree* dan *Random Forest* untuk mengklasifikasikan dan memprediksi kategori karyawan yang akan dipromosikan maupun yang tidak dipromosikan dari data yang dihimpun pada WNS Analytics Wizard 2018: ML Hackathon.

Data yang diolah mengalami ketidakseimbangan atau *imbalanced class*, dimana jumlah kategori yang tidak dipromosikan lebih banyak dibandingkan jumlah kategori yang dipromosikan. Proses klasifikasi dengan *imbalanced class* data akan mengakibatkan buruknya performa dari algoritma klasifikasi yang digunakan. Salah satu cara untuk mengatasi data yang *imbalanced class* adalah dengan menggunakan teknik *oversampling* jenis *Synthetic Minority Over-sampling Technique* (SMOTE) (Kasanah, 2019). Sehingga pada penelitian ini menerapkan teknik *over-sampling* SMOTE beserta algoritma klasifikasi yang telah disebutkan di atas dalam mengklasifikasikan kategori karyawan yang dipromosikan dan kategori karyawan yang tidak dipromosikan.

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

II. METODE PENELITIAN

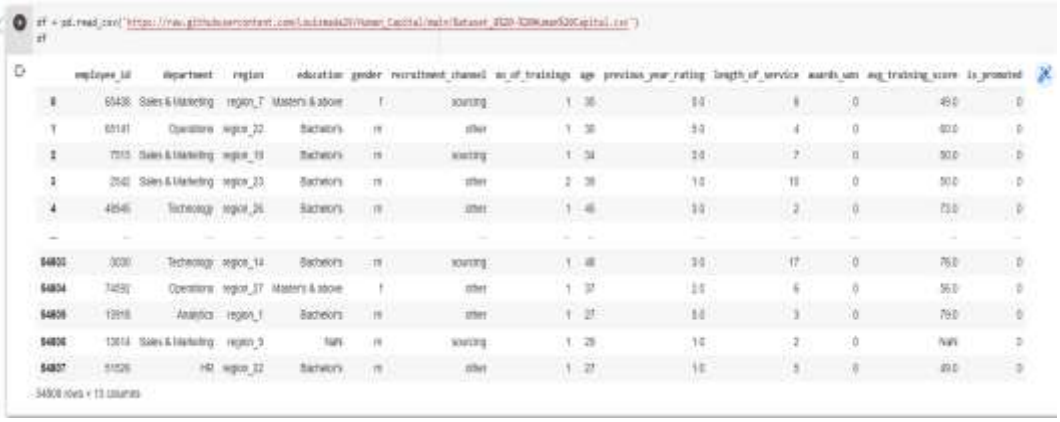
Metodologi penelitian yang dilakukan terdiri dari 4 tahap, yang dapat dilihat pada Gambar 1.



Gambar 1 Metodologi Penelitian

2.1 *Data Understanding*

Pada tahap ini merupakan tahap pengumpulan data, menelaah data untuk memahami data yang akan digunakan, mengidentifikasi masalah dengan memahami substansi dalam data dan mencari hal yang menarik dalam data untuk menemukan hipotesis awal (Budiman, 2012). Data pada penelitian ini diperoleh melalui Kaggle (https://www.kaggle.com/rsnayak/wns-analytics-wizard-2018-ml-hackathon). Dataset yang diperoleh terdiri dari 13 kolom dan 54808 baris.



Gambar 2 Dataset Human Capital WNS Analytics Wizard 2018: ML Hackathon

2.2 *Data Pre-processing*

Proses selanjutnya sebelum model algoritma dibuat adalah data *pre-processing*. Pada penelitian ini dilakukan dilakukan teknik *pre-processing* yang terdiri dari 6 tahap. Tahap pertama adalah *data cleansing*, tahap ini adalah kegiatan

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

untuk menganalisa kualitas data dengan cara melakukan aggregasi data, mengecek *missing value*, melakukan pembersihan data yang terduplikasi dan *data imputation* (Darwis, 2021). Tahap kedua adalah *Label Encoder*, tahap ini adalah kegiatan dengan mengubah label pada data kategorial dengan teknik pengkodean yang memberikan bilangan bulat yang unik berdasarkan urutan abjad (Ardiansyah, 2020). Tahap ketiga adalah *Feature Selection*, tahap ini adalah kegiatan untuk memilih subset fitur dari fitur asli dalam sebuah *dataset* (Somantri, 2017). Pada tahap ini dilakukan dengan metode statistika, yaitu pada data numerik dilakukan *Analysis of Variances* (ANOVA) *Testing* dan pada data kategorial dilakukan *Chi-Squared Testing*. ANOVA *Testing* adalah metode statistika yang dilakukan untuk menganalisis perbandingan multivariabel dengan tujuan untuk menemukan interaksi antar variabel (Hermawan, 2017). Sedangkan *Chi-Squared Testing* adalah metode statistika untuk uji perbandingan non-parametris yang dilakukan pada data nominal (Indriyanto, 2014). Tahap Keempat adalah melakukan *Over-sampling* dengan *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE dipilih untuk menyelesaikan *imbalanced class* pada data dengan cara menyeimbangkan jumlah distribusi pada data kelas minoritas dengan menggunakan data sintesis dari data kelas mayoritas seingga jumlah keduanya seimbang (Kasanah, 2019). Tahap kelima adalah *split data* ke dalam *training set* dan *test set* dengan perbandingan 4:1. Dan tahap ke enam adalah *Standardization*, tahap ini dilakukan dengan menggunakan metode *Standard Scaler*.

2.3 *Develop Model*

Setelah data selesai pada tahap *pre-processing*, selanjutnya adalah membangun model prediksi dengan algoritma *K-Nearest Neighbor* (KNN), *Decision Tree* dan *Random Forest*.

2.3.1 *K-Nearest Neighbor* (KNN)

Algoritma *K-Nearest Neighbor* adalah metode klasifikasi terhadap suatu *dataset* berdasarkan jarak data pembelajaran (*neighbor*) terdekat. Jauh maupun dekatnya data pembelajaran (*neighbor*) tersebut dihitung dengan jarak *Euclidean* (Baharuddin, 2019).

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

2.3.2 *Decision Tree*

Algoritma *Decision Tree* merupakan algoritma yang memiliki struktur menyerupai diagram alir (*flowchart*) yang setiap *node* merupakan pengujian terhadap variabel attribut. Algoritma ini digunakan dalam analisa klasifikasi dan prediksi dalam bentuk pohon keputusan (Sutoyo, 2018).

2.3.3 *Random Forest*

Algoritma ini merupakan salah satu jenis dari algoritma *Decision Tree* yang merupakan kumpulan (*ensemble*) pohon keputusan yang digunakan sebagai *base classifier* yang dikombinasikan. Algoritma ini digunakan untuk analisa klasifikasi dan regresi (Primajaya, 2018).

2.4 *Evaluation Model*

Pada tahap ini model algoritma yang telah diterapkan dalam metode pembelajaran klasifikasi dilakukan perhitungan performa model pada algoritma KNN, *Decision Tree* dan *Random Forest*. Perhitungan performa model klasifikasi didasarkan pada pengujian objek yang benar dan objek yang salah. Perhitungan performa klasifikasi yang digunakan pada penelitian ini adalah *confusion matrix* yang berisi perhitungan hasil klasifikasi aktual yang dapat diprediksi (Kasanah, 2019). Pada Tabel 1 menunjukkan *confusion matrix* dua kelas.

Tabel 1 *Confusion matrix*

	Prediksi Positif	Prediksi Negatif
Aktual Positif	TP	FN
Aktual Negatif	FP	TN

Pada Tabel 1 TP adalah True Positive, TN adalah True Negative, FP adalah *False Positive* dan FN adalah *False Negative*. Pada penelitian ini performa klasifikasi yang akan dihitung adalah *accuracy*, *precision*, *recall* dan F1 *score*. Persamaan (1), (2), (3) dan (4) berturut-turut merupakan rumus perhitungan *accuracy*, *precision*, *recall* dan F1 *score* (Naufal, 2021).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 } Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

III. HASIL DAN PEMBAHASAN

Berdasarkan hasil *Feature Selection* dengan menggunakan metode statistik, maka dapat ditentukan bahwa variabel independen yang digunakan pada penelitian ini, antara lain region, education, no_of_training, age, previous_year_rating, length_of_services, awards_won, dan avg_training_score. Tabel 2 menunjukkan proporsi jumlah data yang digunakan dalam penelitian ini setelah dilakukan penyeimbangan pada kasus *imbalanced class* dengan teknik SMOTE.

Tabel 2 Proporsi Data

Label	*Imbalanced Class*	SMOTE
0	50140	50140
1	4668	50140

3.1 KNN

Pada penelitian ini dalam proses klasifikasi pada algoritma KNN menggunakan nilai K = 7 dengan jarak *Euclidean* yang telah dinormalisasi, tabel 3 menunjukkan hasil evaluasi model klasifikasi dengan algoritma KNN.

Tabel 3 Performa Model KNN

Performa	Training	Testing
TP	37210	9165
TN	33798	8197
FP	6320	1825
FN	2896	869
Accuracy	88.51 %	86.57 %
Precision	92.11 %	90.41 %
Recall	84.25 %	81.79 %
F1 *Score*	88.0 %	85.88 %

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

3.2 *Decision Tree*

Tabel 4 menunjukkan perfoma model klasifikasi dengan algoritma *Decision Tree* dengan *max_depth* = 15 yang digunakan untuk mengontrol ukuran pohon dan mencegah terjadinya *overfitting* dan *criterion* = '*gini'*.

Tabel 4 Performa Model *Decision Tree*

Performa	Training	Testing
TP	36929	9014
TN	33309	8092
FP	6809	1930
FN	3177	1020
Accuracy	87.55 %	85.29 %
Precision	91.29 %	88.81 %
Recall	83.03 %	80.74 %
F1 *Score*	86.96 %	84.58 %

3.3 *Random Forest*

Tabel 5 menunjukkan hasil perfoma model klasifikasi dengan algoritma *Random Forest* dengan *max_depth* = 15 dan *criterion* = '*gini'*.

Tabel 5 Performa Model *Random Forest*

Performa	Training	Testing
TP	37122	9117
TN	33859	8206
FP	6259	1816
FN	2984	917
Accuracy	88.48 %	86.37 %
Precision	91.9 %	89.95 %
Recall	84.4 %	81.88 %
F1 *Score*	87.99 %	85.73 %

Tabel 6 menunjukkan perbandingan performa model algoritma klasifikasi yang digunakan pada data training.

Tabel 6 Perbandingan Performa Model Data Training

Algoritma	*Accuracy*	*Precision*	*Recall*	F1 *Score*
KNN	88.51 %	92.11 %	84.25 %	88.0 %
Decision Tree	87.55 %	91.29 %	83.03 %	86.96 %
Random Forest	88.48 %	91.9 %	84.4 %	87.99 %

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

Tabel 7 menunjukkan perbandingan performa model algoritma klasifikasi yang digunakan pada data testing.

Tabel 7 Perbandingan Performa Model Data Testing

Algoritma	*Accuracy*	*Precision*	*Recall*	F1 *Score*
KNN	86.57 %	90.41 %	81.79 %	85.88 %
Decision Tree	85.29 %	88.81 %	80.74 %	84.58 %
Random Forest	86.37 %	89.95 %	81.88 %	85.73 %

Pada penelitian ini digunakan data yang mengalami *imbalanced class* sehingga F1 *Score* merupakan metriks evaluasi yang lebih tepat untuk mengetahui performa model klasifikasi pada data yang mengalami ketidakseimbangan kelas. Berdasarkan tabel perbandingan performa model diatas menunjukkan bahwa algoritma KNN merupakan algoritma yang terbaik.

IV. KESIMPULAN

Dataset yang digunakan pada penelitian ini merupakan data yang mengalami *imbalanced class*, sehingga untuk menyeimbangkan kelas data diterapkan metode SMOTE. Berdasarkan hasil performa model klasifikasi dari model algoritma yang digunakan pada penelitian ini, maka model KNN memiliki hasil performa yang terbaik nilai metriks evaluasinya. Sehingga pada penelitian ini, model KNN adalah model klasifikasi yang lebih baik digunakan pada penelitian ini, dibandingkan dengan algoritma *Decision Tree* dan *Random Forest*. Pada penelitian ini juga hasil performa model klasifikasi yang dilakukan tidak ditemukan *overfitting* maupun *underfitting*, sehingga model dapat berperforma baik pada *training* maupun *testing*.

DAFTAR PUSTAKA

Aji*Primajaya, B. N. 2018. Random*Forest*Algorithm for Prediction of Precipitation.**Indonesia Journal*of Artificial Intelligence*and Data Mining(IJAIDM)*. vol 1 (1): 27-31.

Anis Nikmatul Kasanah, M. U. 2019. Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objekvitas Berita Online Menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*. vol 3 (2): 196-201.

Ardiansyah, F. 2020. Sistem Prediksi Harga Sewa Kost Dengan Menggunakan Random Forest Analytics (Studi Kasus: Kost Eksklusif di Daerah Istimewa

Yogyakarta) [Tugas Akhir]. Yogyakarta: Universitas Islam Indonesia Yogyakarta.

Citra Ayu Ningsi, T. A. 2015. Pengaruh Pelatihan dan Promosi Terhadap Motivasi dan Kinerja Karyawan (Studi Pada Karyawan PT. PLN (Persero) Area kendari). *Jurnal Administrasi Publiki*. vol *5* (1): 131-143.

Dadi Darwis, N. S. 2021. Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional. *Jurnal Tekno Kompak*. vol 15 (1): 131-145.

Hermawan, D. P. 2017. Efektivitas Penggunaan Game Edukasi Berjenis Puzzle, RPG dan Puzzle RPG Sebagai Sarana Belajar Matematika [Disertasi]. Surabaya: Institut Teknologi Sepuluh November.

Jatmiko Indriyanto, P. C. 2014. Algoritma K-Nearest Neighbor Berbasis Chi-Squared untuk Prediksi Nasabah Asuransi. *Jurnal Dian Nuswantoro University*.

Mardi, Y. 2017. Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Jurnal Edik Informatika*. vol 2 (2): 213-219.

Mus Mulyadi Baharuddin, T. H. 2019. Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca. *ILKOM Jurnal Ilmiah*. vol 11 (3): 269-274.

Naufal, M. F. 2021. Analisis Perbandingan Algoritma SVM, KNN dan CNN untuk Klasifikasi Citra Cuaca. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*. vol 8 (2): 311-317.

Nobertus Krisandi, H. B. 2013. Algoritma K-Nearest Neihbor Dalam Klasifikasi Data Hasil Produksi Kelapa Sawit Pada PT. Minamas Kecamatan Parindu. *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)*. vol 2 (1): 33-38.

Oman Somantri, M. K. 2017. Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naive Bayes dan Algoritme Genetika. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*. vol 6 (3): 301-3016.

Putri, N. K. 2013. Peran Human Capital terhadap Kesuksesan Organisasi: Karyawan Adalah Investasi. *Jurnal Administrasi Kebijakan Kesehatan*. vol 11 (2): 93-95.

Sutoyo, I. 2018. Implementasi Algoritma Decision Tree untuk Klasifikasi Data Peserta Didik. *Jurnal PILAR Nusa Mandiri*. vol 14 (2): 217-224.

Unik Desthiani, S. S. 2018. Peranan Gaya Kepemimpinan, Motivasi dan Disiplin Kerja terhadap Kinerja karyawan. *Jurnal Sekretari Universitas Pamulang*. vol 5 (1): 1-16.