

**MODEL UNTUK MEMPREDIKSI KINERJA PEGAWAI  
DENGAN ALGORITMA C.45**



**TESIS**

Diajukan sebagai salah satu syarat untuk memperoleh gelar Magister  
Ilmu Komputer (M.Kom)

RIZKY ADE SAFITRI  
14002244

Program Studi Ilmu Komputer (S2)  
Sekolah Tinggi Manajemen Informatika dan Komputer  
Nusa Mandiri  
2020

## **SURAT PERNYATAAN ORISINALITAS DAN BEBAS PLAGIARISME**

Yang bertanda tangan di bawah ini :

Nama : Rizky Ade Safitri  
NIM : 14002244  
Program Studi : Ilmu Komputer  
Jenjang : Strata Dua (S2)  
Konsentrasi : Data Mining

Dengan ini menyatakan bahwa tesis yang telah saya buat dengan judul: “Model Untuk Memprediksi Kinerja Pegawai dengan Algoritma C.45 ” adalah hasil karya sendiri, dan semua sumber baik yang kutip maupun yang dirujuk telah saya nyatakan dengan benar dan tesis belum pernah diterbitkan atau dipublikasikan dimanapun dan dalam bentuk apapun.

Demikianlah surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila dikemudian hari ternyata saya memberikan keterangan palsu dan atau ada pihak lain yang mengklaim bahwa tesis yang telah saya buat adalah hasil karya milik seseorang atau badan tertentu, saya bersedia diproses baik secara pidana maupun perdata dan kelulusan saya dari Program Studi Ilmu Komputer (S2) Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri dicabut/dibatalkan

Jakarta, 06 Agustus 2020  
Yang menyatakan,



Rizky Ade Safitri

## HALAMAN PERSETUJUAN DAN PENGESAHAN TESIS

Tesis ini diajukan oleh:

Nama : Rizky Ade Safitri  
NIM : 14002244  
Program Studi : Ilmu Komputer  
Jenjang : Strata Dua (S2)  
Konsentrasi : *Data Mining*  
Judul Tesis : Model Untuk Memprediksi Kinerja Pegawai Dengan Algoritma C.45

Telah dipertahankan pada periode 2020-1 dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh Magister Ilmu Komputer (M.Kom) pada Program Studi Ilmu Komputer (S2) Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri).

Jakarta, 12 Agustus 2020

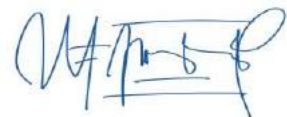
### PEMBIMBING TESIS

Pembimbing I : Dr. Agus Subekti, M.T



### DEWAN PENGUJI

Penguji I : Dr. Hilman Ferdinandus Pardede,  
: S.T, M.EICT



Penguji II : Dr. Lindung Parningotan Manik,  
: M.T.I



Penguji III /  
Pembimbing I : Dr. Agus Subekti, M.T





## LEMBAR BIMBINGAN TESIS

STMIK Nusa Mandiri

NIM : 14002244  
Nama Lengkap : Rizky Ade Safitri  
Dosen Pembimbing : Dr. Agus Subekti, M.T  
Judul Tesis : **“Model Untuk Memprediksi Kinerja Pegawai dengan Algoritma C.45”**

No	Tanggal Bimbingan	Materi Bimbingan	Paraf Dosen Pembimbing
1.	02 April 2020	Pengajuan Proposal Tesis	
2.	11 April 2020	Pengajuan BAB I	
3.	25 April 2020	ACC Pengajuan BAB I dan Pengajuan BAB II	
4.	09 Mei 2020	ACC Pengajuan BAB II dan Pengajuan BAB III	
5.	18 Mei 2020	ACC Pengajuan BAB III dan Pengajuan BAB IV	
6.	31 Mei 2020	Pengujian dan Pembahasan BAB V	
7.	27 Juli 2020	ACC Pengajuan BAB IV dan Pengajuan BAB V	
8.	05 Agustus 2020	ACC Keseluruhan	

Catatan untuk Dosen Pembimbing  
Bimbingan Tesis

- Dimulai pada tanggal : 02 April 2020
- Diakhiri pada tanggal : 05 Agustus 2020
- Jumlah pertemuan bimbingan : 8 (delapan) Kali

Disetujui Oleh,  
Dosen Pembimbing

**Dr. Agus Subekti, M.T**

## KATA PENGANTAR

Puji syukur alhamdulillah, penulis panjatkan kehadiran Allah SWT, yang telah melimpahkan rahmat dan karunia-Nya, sehingga pada akhirnya penulis dapat menyelesaikan tesis ini tepat pada waktunya. Dimana tesis ini penulis sajikan dalam bentuk buku yang sederhana. Adapun judul tesis, yang penulis ambil sebagai berikut **“Model Untuk Memprediksi Kinerja Pegawai dengan Algoritma C.45 ”**.

Tujuan penulisan tesis ini dibuat sebagai salah satu syarat untuk mendapatkan gelar Magister Ilmu Komputer (M.Kom) pada Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri).

Penulis menyadari bahwa tanpa bimbingan dan dukungan dari semua pihak dalam pembuatan tesis ini, maka penulis tidak dapat menyelesaikan tesis ini tepat pada waktunya. Untuk itu ijinilah penulis pada kesempatan ini untuk mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Ibu Dr. Dwiza Riana, S.SI, MM, M.Kom selaku Ketua Pasca Sarjana Magister Ilmu Komputer STMIK Nusa Mandiri.
2. Bapak Dr. Agus Subekti, M.T yang sudah dengan setulus hati membimbing, mengarahkan menyumbangkan ide, waktu, dan tenaganya dalam membimbing penulis untuk menyelesaikan tesis ini.
3. Orang tua tercinta yang telah memberikan dukungan doa, material dan moral kepada penulis.
4. Seluruh staf pengajar (dosen) STMIK Nusa Mandiri yang telah memberikan pelajaran yang berarti bagi penulis selama menempuh studi.
5. Seluruh sahabat dan teman-teman seperjuangan yang selalu mendukung dan membantu.

Serta semua pihak yang terlalu banyak untuk penulis sebutkan satu persatu sehingga terwujudnya penulisan tesis ini. Penulis menyadari bahwa penulisan ini masih jauh sekali dari sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan dimasa yang akan datang.

Akhir kata semoga tesis ini dapat berguna bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Jakarta, 06 Agustus 2020

Penulis

A handwritten signature in black ink, appearing to read 'Rizky' followed by a stylized surname.

Rizky Ade Safitri

## SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan di bawah ini, saya :

Nama : Rizky Ade Safitri  
NIM : 14002244  
Program Studi : Ilmu Komputer  
Jenjang : Strata Dua (S2)  
Konsentrasi : Data Mining  
Jenis Karya : Tesis

Demi pengembangan ilmu pengetahuan, dengan ini menyetujui untuk memberikan ijin kepada pihak Program Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri) **Hak Bebas Royalti Non-Eksklusif (*Non-exclusive Royalti-Free Right*)** atas karya ilmiah kami yang berjudul : “Model Untuk Memprediksi Kinerja Pegawai dengan Algoritma C.45” beserta perangkat yang diperlukan (apabila ada).

Dengan **Hak Bebas Royalti Non-Eksklusif** ini pihak STMIK Nusa Mandiri berhak menyimpan, mengalih-media atau *bentuk-kan*, mengelolanya dalam pangkalan data (*database*), mendistribusikannya dan menampilkan atau mempublikasikannya di *internet* atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari kami selama tetap mencantumkan nama kami sebagai penulis atau pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak STMIK Nusa Mandiri, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 06 Agustus 2020

Yang menyatakan,



Rizky Ade Safitri

## ABSTRAK

Nama : Rizky Ade Safitri  
NIM : 14002244  
Program Studi : Magister Ilmu Komputer  
Jenjang : Strata Dua (S2)  
Konsentrasi : *Data Mining*  
Judul : “Model Untuk Memprediksi Kinerja Pegawai dengan Algoritma C.45 ”

Manajemen Sumber Daya Manusia (SDM) menjadi salah satu kepentingan esensial manajer dan pengambil keputusan di hampir semua jenis bisnis untuk diadopsi rencana untuk menemukan karyawan yang berkualifikasi dengan benar. HRM bertanggung jawab mengalokasikan karyawan terbaik untuk yang sesuai pekerjaan pada waktu yang tepat, latih dan kualifikasi mereka, dan bangun sistem evaluasi untuk memantau kinerja mereka dan upaya untuk melestarikan bakat potensial karyawan. Dengan klasifikasi, model Prediktif miliki target spesifik yang memungkinkan kami memprediksi hal yang tidak diketahui nilai variabel tergantung pada minat sebelumnya nilai yang diketahui dari variabel lain. Dalam penelitian ini menjelaskan teknik resample dengan penambahan fitur selection algoritma yaitu correlation attribute eval. Dari hasil pengujian klasifikasi menggunakan teknik resample pada dataset HRM menggunakan algoritma *Desicion Tree J48*, *Random Forest*, *Naive Bayes*, *KNN*, *Logistic* dan *SVM* menunjukkan hasil akurasi terbaik yaitu algoritma Decision Tree J48 dengan nilai akurasi sebesar 95,41%, nilai kappa sebesar 0,8925, nilai MAE sebesar 0,0432, Nilai Precision sebesar 0,955, Nilai Recall sebesar 0,954 dan nilai ROC sebesar 0,964.

Kata kunci: Algoritma *C45(J48)*, *Resample*, *Kinerja*, Klasifikasi



## **ABSTRACT**

Name : Rizky Ade Safitri  
NIM : 14002244  
Study program : Magister Ilmu Komputer  
Jenjang : Strata Dua (S2)  
Concentration : *Data Mining*  
Title : *“Model for Predicting Employee Performance with C.45 Algorithm”*

*Human Resource (HR) management is one of the essential interests of managers and decision makers in almost any type of business to adopt a plan to find properly qualified employees. HRM is responsible for allocating employees to the appropriate job at the right time, training and qualifying them, and establishing an evaluation system to assess their performance and efforts to preserve the potential talents of employees. By classification, the Predictive model that has a specific target that we can estimate is not in accordance with the predetermined value. In this research, it explains the resample technique with the addition of the algorithm selection feature, namely the evaluation attribute correlation. From the results of the classification examiners using the resample technique on the HRM dataset using the Decision Tree J48 algorithm, Random Forest, Naive Bayes, KNN, Logistics and SVM shows the best results, namely the J48 Decision Tree algorithm with a value of 95.41%, a kappa value of 0.8925. , MAE value 0.0432, precision value 0.955, recall value 0.954 and ROC value 0.964.*

*Keywords: C45 Algorithm (J48), Resample, Performance, Classification*

## DAFTAR ISI

	Halaman
HALAMAN JUDUL .....	i
SURAT PERNYATAAN ORISINALITAS DAN BEBAS PLAGIARISME .....	ii
PERSETUJUAN TESIS .....	iii
PERSETUJUAN DAN PENGESAHAN .....	iv
LEMBAR BIMBINGAN TESIS .....	v
KATA PENGANTAR .....	<b>Error! Bookmark not defined.</b>
SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS .....	<b>Error! Bookmark not defined.</b>
ABSTRAK .....	ix
<i>ABSTRACT</i> .....	<b>Error! Bookmark not defined.</b>
DAFTAR ISI .....	xi
DAFTAR TABEL .....	xii
DAFTAR GAMBAR .....	xiii
BAB I PENDAHULUAN .....	xiv
1.1. Latar Belakang Penulisan .....	1
1.2. Identifikasi Masalah .....	2
1.3. Tujuan Penelitian .....	3
1.4. Ruang Lingkup Penelitian .....	3
1.5. Sistematika Penulisan .....	3
BAB II LANDASAN TEORI .....	6
2.1. Tinjauan Pustaka .....	6
2.1.1. Decision Tree C.45 .....	6
2.1.2. Naive Bayes .....	8
2.1.3. Support Vector Machine .....	8
2.1.4. K-Nearest Neighbor .....	8
2.1.5. Random Forest .....	8

2.1.6.	K-Fold Cross Validation .....	9
2.1.7.	Preprocessing Data.....	9
2.1.8.	Confusion Matrix .....	9
2.1.9.	Klasifikasi .....	11
2.1.10.	Dataset.....	11
2.1.11.	Teknik Resample.....	12
2.1.12.	Data Mining .....	13
2.2.	Tinjauan Penelitian Terkait .....	13
BAB III METODOLOGI PENELITIAN.....		16
3.1.	Metodologi Penelitian .....	16
3.1.1.	Jenis Penelitian.....	17
3.1.2.	Metode Pengumpulan Data .....	18
BAB IV HASIL DAN PEMBAHASAN .....		21
4.1.	Dataset .....	21
4.2.	Preprocessing Data .....	22
4.2.1	Preprocessing Data Sebelum di Resample.....	24
4.2.2	Proses Preprocessing Data di Resample.....	25
4.2.3	Preprocessing Data Sesudah di Resample .....	27
4.3.	Eksperimen Dan penngujian Model .....	27
4.3.1	Pengujian Model Klasifikasi .....	28
4.3.2	Pengujian Model Klasifikasi Dataset Resample .....	31
4.3.3	Pengujian Klasifikasi dengan resample dan feature selection .....	33
BAB V PENUTUP.....		39
5.1.	Kesimpulan.....	39
5.2.	Saran .....	40
DAFTAR REFERENSI .....		41
DAFTAR RIWAYAT HIDUP.....		44

## DAFTAR TABEL

	Halaman
Table II.1 Penelitian Terkait .....	<b>Error! Bookmark not defined.</b>
Tabel IV.1 Spesifikasi Dataset.....	21
Tabel IV.2 Penjelasan Atribut pada <i>dataset</i> HRM.....	21
Tabel IV.3 Hasil <i>Confusion Matrix</i> Algoritma <i>Decision Tree J48</i> .....	29
Tabel IV.4 Hasil <i>Confusion Matrix</i> Algoritma <i>Random Forest</i> .....	29
Tabel IV.5 Hasil <i>Confusion Matrix</i> Algoritma <i>Naive Bayes</i> .....	29
Tabel IV.6 Hasil <i>Confusion Matrix</i> Algoritma <i>Support Vector Machine (SVM)</i> .	29
Tabel IV.7 Hasil <i>Confusion Matrix</i> Algoritma <i>Logistic</i> .....	30
Tabel IV.8 Hasil <i>Confusion Matrix</i> Algoritma <i>K-Nearest Neighbor</i> .....	30
Tabel IV.9 Hasil Kinerja Klasifikasi dari algoritma <i>Decision Tree J48, Random Forest, SVM, Naive Bayes, Logistic, KNN</i> .....	31
Tabel IV.10 Hasil <i>Confusion Matrix</i> Algoritma <i>Decision Tree J48</i> dengan <i>Resample</i> .....	31
Tabel IV.11 Hasil <i>Confusion Matrix</i> Algoritma <i>Random Forest</i> dengan <i>Resample</i> .....	32
Tabel IV.12 Hasil <i>Confusion Matrix</i> Algoritma <i>Naive Bayes</i> dengan <i>Resample</i> .	32
Tabel IV.13 Hasil <i>Confusion Matrix</i> Algoritma <i>SVM</i> dengan <i>Resample</i> .....	32
Tabel IV.14 Hasil <i>Confusion Matrix</i> Algoritma <i>Logistic</i> dengan <i>Resample</i> .....	33
Tabel IV.15 Hasil <i>Confusion Matrix</i> Algoritma <i>K-Nearest Neighbor</i> dengan <i>Resample</i> .....	33
Tabel IV.16 Hasil Kinerja Pengujian Klasifikasi dengan <i>Resample</i> .....	33
Tabel IV.17 Rank Attributes .....	36

## DAFTAR GAMBAR

	Halaman
Gambar II.1 Klasifikasi dan Prediksi .....	10
Gambar III.1 Metodologi Penelitian .....	16
Gambar III.2 Sampel Dataset .....	19
Gambar III.3 Sampel Dataset Lanjutan 1 .....	19
Gambar III.4 Sampel Dataset Lanjutan 2 .....	19
Gambar IV.1 Tampilan awal weka .....	23
Gambar IV.2 Tampilan Antarmuka weka .....	23
Gambar IV.3 Tampilan Dataset di weka .....	24
Gambar IV.4 Tampilan Preprocessing Data Sebelum di Resample .....	25
Gambar IV.5 Tampilan Proses Preprocessing Data di Resample .....	26
Gambar IV.6 Tampilan <i>weka.gui.GenericObjectEditor</i> .....	26
Gambar IV.7 Tampilan Preprocessing Data Sesudah di Resample .....	27
Gambar IV.8 Alur Metode yang diusulkan .....	28
Gambar IV.9 Grafik Hasil Kinerja .....	34
Gambar IV.10 Visualize All Atribut Dataset Asli Tanpa Resample .....	35
Gambar IV.11 Visualize All Atribut Dataset Asli + Resample .....	36

# **BAB I**

## **PENDAHULUAN**

### **1.1. Latar Belakang Penulisan**

Sumber daya manusia merupakan investasi sangat berharga bagi sebuah organisasi yang perlu dijaga. Manfaat dari adanya pengembangan SDM itu sendiri yaitu untuk peningkatan produktifitas kerja, terwujudnya hubungan yang serasi antara atasan dan bawahan, tersedianya proses pengambilan keputusan yang cepat dan tepat serta meningkatnya semangat kerja seluruh anggota dalam organisasi. Manajemen Sumber Daya Manusia (SDM) menjadi salah satu kepentingan esensial manajer dan pengambil keputusan di hampir semua jenis bisnis untuk diadopsi rencana untuk menemukan karyawan yang berkualifikasi dengan benar.

HRM memiliki peran utama dalam memutuskan daya saing dan efektivitas menjadi lebih baik kelanjutan. Organisasi menganggap HRM sebagai “orang praktik”. Oleh karena itu, menjadi tanggung jawab HRM mengalokasikan karyawan terbaik untuk yang sesuai pekerjaan pada waktu yang tepat, latih dan kualifikasi mereka, dan bangun sistem evaluasi untuk memantau kinerja mereka dan upaya untuk melestarikan bakat potensial karyawan. Teknik klasifikasi yang digunakan umumnya membangun model, yang pada gilirannya digunakan untuk memprediksi data masa depan tren. Dengan klasifikasi, model Prediktif memiliki target spesifik yang memungkinkan kami memprediksi hal yang tidak diketahui nilai variabel tergantung pada minat sebelumnya nilai yang diketahui dari variabel lain. Untuk meningkatkan kemajuan dalam keberhasilan suatu perusahaan juga diperlukan kinerja yang baik dan berkualitas yang sangat tinggi dari para pegawai suatu perusahaan tersebut.

Maka dari itu peneliti melakukan penelitian model untuk memprediksi peringkat kinerja pegawai. Dimana dalam memprediksi peringkat kinerja pegawai tersebut menggunakan beberapa algoritma yaitu *Decision Tree J48*, *Random*

*Forest, Naive Bayes, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbord (KNN)*. Kemudian ditambah dengan *Feature selection algorithm* yaitu *correlation attribute eval* untuk mengetahui variabel apa saja yang menjadi faktor pengaruh dalam prediksi peringkat kinerja pegawai.

Penelitian sebelumnya tentang kinerja pegawai menggunakan algoritma *Decision Tree C.45, SVM*, dan *Naive Bayes* melakukan penelitian dengan menggunakan *fitur selection algorithm* yaitu *Correlation-Attributeeval* , *Gainratio-Attributeeval* , *Relieff-Attributeeval* menunjukkan bahwa ketiga teknik memiliki konvergen dan akurasi sedang, yang lebih besar dari 70%. Itu akurasi moderat dapat dianggap dapat diterima akurasi dalam banyak kasus. Dalam ketiga percobaan, dataset menghasilkan model yang memuaskan untuk masing-masing dari ketiganya teknik klasifikasi yang dipilih. SVM Teknik ditemukan sebagai classifier yang paling cocok untuk membangun model prediksi, di mana ia memiliki yang terbesar akurasi prediksi melalui ketiga percobaan yang telah dieksekusi dengan persentase tertinggi 86,90%.

Berdasarkan paparan di atas penulis akan menggunakan data yang lebih banyak dari paper rujukan yaitu berjumlah 1200 data juga menambahkan teknik resample untuk menyeimbangkan class minoritas dan mayoritas supaya mendapatkan hasil akurasi yang lebih baik. Model yang dihasilkan diharapkan bisa membantu dalam memprediksi peringkat kinerja pegawai juga mengetahui variabel-variabel mana yang sangat berpengaruh untuk peringkat kinerja pegawai. Dengan ini dibuatlah tesis berjudul “**Model Untuk Memprediksi Kinerja Pegawai dengan Algoritma C.45**”.

## 1.2. Identifikasi Masalah

Beberapa permasalahan yang bisa diuraikan sebagai berikut :

1. Bagaimana melakukan praprocessing pada dataset HRM?
2. Model algoritma *Decision Tree J.48, Random Forest, Naive Bayes, Support Vector Machine, K-Nearest Neighbor*, dan *Logistic* manakah yang menghasilkan hasil prediksi paling akurat dalam memprediksi klasifikasi dataset HRM ?

3. Bagaimana hasil dari model klasifikasi dengan algoritma *Decision Tree J.48*, *Random Forest*, *Naive Bayes*, *Support Vector Machine*, *K-Nearest Neighbor*, dan *Logistic* menggunakan teknik *resample* manakah kinerja yang dapat menghasilkan nilai akurasi yang terbaik ?
4. Dari penggunaan *feature selection algorithm* yaitu *correlation attribute eval* atribut apa sajakah yang paling berpengaruh dalam memprediksi peringkat kinerja karyawan pada dataset HRM ?

### 1.3. Tujuan Penelitian

Tujuan dari penelitian ini melakukan prediksi peringkat kinerja karyawan dengan menggunakan model teknik *resample* juga *Feature Selection Algorithm* yaitu *Correlation Attribute Eval* untuk memperoleh hasil akurasi tertinggi dan atribut yang berpengaruh dalam peringkat kinerja karyawan. Penelitian ini diharapkan dapat menjadi pembanding dengan penelitian-penelitian sejenis dan terkait.

### 1.4. Ruang Lingkup Penelitian

Ruang lingkup ini berfungsi untuk membatasi pembahasan pada pokok permasalahan yaitu pengujian klasifikasi dengan model teknik *resample* dan *feature selection algorithm* yaitu *correlation attribute eval* dengan menggunakan algoritma *Decision Tree J.48*, *Random Forest*, *Naive Bayes*, *Support Vector Machine*, *K-Nearest Neighbor*, dan *Logistic*.

### 1.5. Sistematika Penulisan

Sistematika penulisan tesis ini terdiri dari lima bab, dimana setiap bab-nya terdiri dari sub bab sebagai berikut :



## BAB I PENDAHULUAN

Pendahuluan pada Bab I membahas tentang Latar Belakang Penulisan, Identifikasi Masalah, Tujuan Penelitian, Hipotesis dan Sistematika Penulisan.

## BAB II LANDASAN/KERANGKA PEMIKIRAN

Landasan atau Kerangka Pemikiran pada Bab II membahas tentang teori yang melandasi penelitian, dalam bab ini juga diuraikan Tinjauan Pustaka, Tinjauan Studi, Tinjauan Organisasi atau Objek Penelitian.

## BAB III METODOLOGI PENELITIAN

Metodologi Penelitian pada Bab III membahas tentang metode pengumpulan data dan eksperimen. Eksperimen merupakan inti dari pembahasan dari bab ini, yaitu menguji dan mengklasifikasi dengan model teknik *resample* dan *feature selection algorithm* yaitu *correlation attribute eval*.

## BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Hasil penelitian dan Pembahasan pada Bab IV membahas tentang hasil dari eksperimen, baik sebelum diterapkan model maupun setelah diterapkan model. Membahas metode, mengukur hasil dengan metode statistik. Hasil pengujian model tersebut dibandingkan untuk melihat tingkat akurasi yang tertinggi. Hasil pengujian model akan ditampilkan dalam bentuk grafik.

## BAB V PENUTUP

Penutup pada Bab V membahas tentang kesimpulan dari penelitian, kekurangan serta kelebihan dari model yang digunakan.

## **BAB II**

### **LANDASAN/KERANGKA PEMIKIRAN**

#### **2.1. Tinjauan Pustaka**

Dalam penelitian ini, penulis menggunakan berbagai referensi yang diambil dari beberapa jurnal Nasional maupun Internasional, Prosiding, website dan buku guna untuk mendukung penelitian ini.

##### **2.1.1. Decision Tree C.45**

Algoritma C.45 merupakan algoritma pengklasifikasi dengan teknik pohon keputusan yang terkenal dan disukai karena memiliki kelebihan-kelebihan. Kelebihan ini misalnya dapat mengolah data numerik (kontinu) dan diskrit, dapat menangani nilai atribut yang hilang, menghasilkan aturan aturan yang mudah diinterpretasikan dan tercepat diantara algoritma-algoritma yang lain. Keakuratan prediksi yaitu kemampuan model untuk dapat memprediksi label kelas terhadap data baru atau yang belum diketahui sebelumnya dengan baik. Dalam hal kecepatan atau efisiensi waktu komputasi yang diperlukan untuk membuat dan menggunakan model. Kemampuan model untuk memprediksi dengan benar walaupun data ada nilai dari atribut yang hilang, dan juga skalabilitas yaitu kemampuan untuk membangun model secara efisien untuk data berjumlah besar (aspek ini akanmendapatkan penekanan)[1].

Proses klasifikasi terdiri dari dua tahap,yaitu tahap belajar dari data pelatihan untuk menghasilkan model dan tahap klasifikasi yang menggunakan model untuk prediksi kelas. Pada tahap belajar dari data, algoritma C4.5 Mengkonstruksi decison tree dari data pelatihan, yang berupa kasus-kasus atau record-record. Setiap kasus berisikan nilai dari atribut-atribut untuk sebuah kelas. Setiap atribut dapat berisi data diskret atau kontinyu(numerik). C4.5 juga menangani kasus yang tidak memiliki nilai untuk sebuah atau lebih atribut. Akan tetapi, atribut kelas hanya bertipe nominal dan tidak boleh kosong[2].

Komponen-komponen yang menyusun algoritma C.45 dalam bentuk pohon keputusan yaitu:

a. Entropy

Entropy merupakan distribusi probabilitas dalam teori informasi dan diadopsi ke dalam algoritma C4.5 untuk mengukur tingkat homogenitas distribusi kelas dari sebuah himpunan (*data set*). Sebagai ilustrasi semakin tinggi tingkat entropy dari sebuah data set maka semakin homogen distribusi kelas pada data set tersebut. Perhitungan entropy ditunjukkan pada persamaan (1).

$$Entropy(S) = -\sum_{i=1}^n p_i \log_2 p_i$$

Dimana  $S$  adalah himpunan kasus,  $n$  adalah jumlah partisi  $S$ , dan  $P_i$  adalah proporsi dari  $S_i$  terhadap  $S$ .

b. Information Gain

Setelah membagi *data set* berdasarkan sebuah atribut ke dalam subset yang lebih kecil, entropy dari data tersebut akan berubah. Perubahan entropy ini dapat digunakan untuk menentukan bagus tidaknya pembagian data yang telah dilakukan. Perubahan entropy ini disebut dengan *information gain* dalam algoritma C4.5. *Information gain* ini diukur dengan mengitung selisih antara entropy data set sebelum dan sesudah pembagian (*splitting*) dilakukan. Pembagian yang terbaik akan menghasilkan entropy subset yang paling kecil, dengan demikian berdampak pada *information gain* yang terbesar. Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada dan dapat ditunjukkan dengan persamaan (2).

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} Entropy(S_i) \quad (2)$$

Dimana  $S$  adalah himpunan kasus,  $A$  adalah atribut yang dihitung,  $n$  adalah jumlah partisi pada atribut  $A$ ,  $|S_i|$  adalah jumlah kasus pada partisi ke  $i$ , dan  $|S|$  adalah jumlah kasus dalam  $S$  [3].

### **2.1.2. Naive Bayes**

Algoritma Naive Bayes merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan Naive dimana diasumsikan kondisi antar atribut saling bebas[4].

### **2.1.3. Support Vector Machine**

Support Vector Machine adalah suatu algoritma yang dikembangkan oleh Boser, Guyon, Vapnik. Algoritma ini untuk pertama kalinya dipresentasikan ke publik pada tahun 1992 di Annual Workshop on Computational Learning Theory. Konsep dasar dari Support Vector Machine merupakan kombinasi dari teori-teori komputasi yang telah ada sebelumnya. Prinsip dasar SVM adalah linear classifier yang berarti hanya bisa digunakan untuk mengklasifikasi data antara 2 kelas namun karena kasus pada dunia nyata umumnya adalah lebih dari 2 kelas maka dikembangkan lebih lanjut agar dapat bekerja pada masalah non-linear atau data non-linear dengan memasukkan konsep kernel trick dan menggunakan fungsi  $\Phi$  agar dapat memetakan data kedalam ruang berdimensi tinggi[5].

### **2.1.4. K-Nearest Neighbor**

K-Nearest Neighbor merupakan salah satu algoritma pembelajaran mesin sederhana. Hal ini hanya didasarkan pada gagasan bahwa suatu objek yang ‘dekat’ satu sama lain juga akan memiliki karakteristik yang mirip. Ini berarti jika kita mengetahui ciri-ciri dari salah satu objek, maka kita juga dapat memprediksi objek lain berdasarkan tetangga terdekatnya. K-NN adalah improvisasi lanjutan dari teknik klasifikasi Nearest Neighbor[6].

### **2.1.5. Random Forest**

Operator ini menghasilkan satu set sejumlah tertentu pohon random yaitu menghasilkan forest (hutan; kumpulan pohon) acak. Model yang dihasilkan

adalah model suara pilihan dari semua pohon. Operator *Random Forest* menghasilkan satu set pohon acak. Pohon-pohon acak yang dihasilkan dengan cara yang persis sama seperti operator Acak Pohon menghasilkan pohon. Model hutan yang dihasilkan mengandung sejumlah tertentu dari model pohon acak. Jumlah pohon parameter menentukan jumlah yang diperlukan pohon. Model yang dihasilkan adalah model suara pilihan dari semua pohon acak. Untuk informasi lebih lanjut tentang pohon acak silakan mempelajari operator *random Tree*[7][8].

#### **2.1.6. K-Fold Cross Validation**

*K-Fold Cross Validation* digunakan karena merupakan salah satu metode yang terbaik untuk memvalidasi data yang akan digunakan. Salah satu contoh dari *K-Fold Cross Validation* adalah *10-Fold Cross Validation*. Teknik ini akan membagi kumpulan data menjadi 10 subset dengan ukuran yang sama, sembilan dari 10 subset data digunakan untuk pelatihan, sementara satu subset yang tertinggal digunakan untuk pengujian. Proses diulang selama sepuluh kali, dan hasil akhirnya diperkirakan sebagai tingkat kesalahan rata-rata pada contoh uji[9].

#### **2.1.7. Preprocessing Data**

*Pre-processing* data adalah proses mengubah data ke dalam format yang sederhana, lebih efektif, dan sesuai dengan kebutuhan pengguna. Indikator yang dapat digunakan sebagai referensi adalah hasil lebih akurat, waktu komputasi yang lebih pendek, juga data menjadi lebih kecil tanpa mengubah informasi di dalamnya[8].

Pra-pemrosesan data adalah salah satu langkah terpenting dalam pembelajaran mesin, yang membantu dalam membangun model pembelajaran mesin dengan lebih akurat. Setiap data scientist harus menghabiskan 80% waktu untuk pra-pemrosesan data dan 20% waktu untuk benar-benar melakukan analisis. Pra-pemrosesan data adalah proses pembersihan data mentah. Data dikumpulkan dalam data raw dan dikonversi ke set data bersih. Dengan kata lain, setiap kali data dikumpulkan dari sumber yang berbeda dikumpulkan dalam format raw dan data ini tidak layak untuk analisis. Oleh karena itu, langkahlangkah tertentu

dijalankan untuk mengubah data menjadi kumpulan data kecil yang bersih, bagian dari proses ini disebut data pra-pemrosesan[13].

### 2.1.8. Confusion Matrix

Confusion Matrix adalah tools yang digunakan untuk evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah matrix dari prediksi yang akan dibandingkan dengan kelas yang asli dari inputan atau dengan kata lain berisi informasi nilai actual dan prediksi pada klasifikasi.

<i>Classification</i>	<i>Predicted class</i>	
	<i>Class = Yes</i>	<i>Class = No</i>
<i>Class = Yes</i>	<i>a (true positive-TP)</i>	<i>b (false negative-FN)</i>
<i>Class = No</i>	<i>c (false positive-FP)</i>	<i>d (true negative-TN)</i>

Gambar II.1 Klasifikasi dan Prediksi

Evaluasi dan validasi hasil dihitung menggunakan rumus akurasi, precision recall berikut ini :

a. Akurasi

Perhitungan akurasi dilakukan dengan cara membagi jumlah data yang diklasifikasi secara benar dengan total sample data testing yang diuji.

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

b. Precision

Menghitung nilai precision dengan cara membagi jumlah data benar yang bernilai positif (True Positive) dibagi dengan jumlah data benar yang bernilai positif (True Positive) dan data salah yang bernilai positif (False Negative).

$$Precision = \frac{TP}{TP+FP}$$

c. Recall

Sedangkan recall dihitung dengan cara membagi data benar yang bernilai positive (True Positive) dengan hasil penjumlahan dari data benar yang bernilai positif (True Positive) dan data salah yang bernilai negatif (False Negative). [10]

$$Recall = \frac{TP}{TP+FN}$$

(5)

### 2.1.9. Klasifikasi

Klasifikasi merupakan proses menemukan model atau fungsi yang menggambarkan, membedakan *class* data atau konsep dengan tujuan agar bisa digunakan untuk *class prediction* dari objek yang *label class* tidak diketahui. Klasifikasi banyak digunakan untuk mendeteksi *fraud* atau penipuan, target pemasaran, prediksi kinerja, manufaktur, dan mendiagnosa kesehatan. Tahapan klasifikasi data terdiri dari 2 langkah. Pertama yaitu tahap *learning* atau *fase* pembelajaran, dimana algoritma klasifikasi untuk menganalisis data *training* atau data latih lalu direpresentasikan ke bentuk model klasifikasi. Kedua adalah proses tahapan klasifikasi, dimana *datatesting* digunakan untuk memprediksi nilai *accuracy* dari model klasifikasi. Jika nilai akurasi *acceptable* atau dapat diterima, maka *rule* bisa diterapkan pada klasifikasi *tupel* data baru[11].

### 2.1.10. Dataset

Dataset adalah kumpulan data yang berelasi/ berkaitan satu dengan lainnya dalam satu kesatuan yang biasanya bersifat spesifik terhadap suatu kasus tertentu, misalnya dataset medikal, dataset komentar pengguna Twitter terhadap layanan Apple, dataset curah hujan selama 1 tahun, dataset pergerakan harga emas selama tahun tertentu, dan lain sebagainya. Dataset bidang medis bahkan biasanya bersifat sangat banyak, kompleks, heterogen, dan hierarkis (Hosseinkhah,



Ashktorab, Veen, & Owrang, 2009). Dataset dapat direpresentasikan dalam berbagai bentuk misalnya bentuk tabel dalam basis data, bentuk matriks, bentuk teks, bentuk *Comma Separated Value* (CSV) dan sebagainya. Dataset dapat dikumpulkan dan dibentuk oleh seseorang, sekelompok atau bahkan suatu organisasi dan dipublikasikan secara online seperti misalnya situs DataHub ([www.datahub.io](http://www.datahub.io)) yang berisi berbagai macam dataset[12].

Dataset dipergunakan sebagai referensi data yang valid untuk suatu penelitian selanjutnya, misalnya untuk referensi data dalam pembelajaran sistem cerdas (sistem pengenalan pola, machine learning, dan lain-lain), atau juga sebagai referensi data dalam pengujian sistem otomatis seperti misalnya pada sistem klasifikasi, klasterisasi, dan sentimen analisis. Dataset yang baik memiliki ciri memiliki data yang lengkap, selalu *up to date*, bersifat konsisten dalam representasi datanya, jumlah variabelnya jelas, tidak mengandung *noise*, menarik, dan mudah dimengerti (Hosseinkhah, Ashktorab, Veen, & Owrang, 2009)[12].

#### **2.1.11. Teknik Resample**

Teknik resampling adalah preprocessing yang mana menyeimbangkan distribusi data kembali untuk mengurangi efek distribusi kelas yang tidak seimbang dalam proses data training. Teknik resampling digunakan untuk mengatasi masalah data tidak seimbang. Metode ini menyeimbangkan data yang asli berdasarkan serangkaian metode algoritma sampling menyesuaikan jumlah sampel kelas yang berbeda, kemudian melatih data seimbang baru yang mengadopsi algoritma klasifikasi[11].

Metode *resampling* terbagi menjadi tiga kategori yaitu Metode *OverSampling*, Metode *UnderSampling*, dan *Hybrid* yang menggabungkan kedua Metode *Sampling*. Teknik *resampling* yang digunakan yaitu *UnderSampling* yang secara acak memilih sampel kelas mayoritas dan menambahkannya ke dalam kelas minoritas sehingga membentuk sebuah *datasets training* baru. Tujuan dari Metode *OverSampling* ini untuk meningkatkan sampel kelas minoritas sampai *equal* atau setara dengan kelas mayoritas lain yang menduplikasi secara acak sampel kelas minoritas[11].

Sedangkan Metode *UnderSampling* menghasilkan *subsampel* acak dari *instance* kelas mayoritas. Metode *UnderSampling* secara acak memilih sampel kelas mayoritas dan menambahkannya ke kelas minoritas sehingga membentuk sebuah *datasets training* baru (1).[11]

### 2.1.12. Data Mining

*Data Mining* dan Analisis Kebutuhan Sistem memiliki suatu keterkaitan antara satu dengan yang lainnya sesuai dengan bidangnya masing-masing. Untuk itu pemanfaatan teknologi dan sumber daya yang ada merupakan salah satu faktor yang perlu diperhatikan bagi seseorang yang akan melakukan pengolahan data. Dalam perkembangannya *data mining* memiliki banyak definisi yang cukup beragam sehingga *data mining* dapat menambah ilmu pengetahuan. *Data mining* adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basis data[4].

Data yang akan diproses berupa data yang sangat besar. Tujuan *data mining* adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat[4].

## 2.2. Tinjauan Penelitian Terkait

Tinjauan penelitian terkait dalam penelitian ini adalah sebagai berikut :

Tabel II.1 Penelitian Terkait

No	Paper	Uraian
1	A proposed Model for Predicting Employees' Performance Using Data Mining Techniques: Egyptian Case Study[14]	Manajemen Sumber Daya Manusia (SDM) telah menjadi salah satu kepentingan penting manajer dan pengambil keputusan di hampir semua jenis bisnis untuk mengadopsi rencana untuk menemukan karyawan yang berkualitas tinggi dengan benar. Oleh karena itu, manajemen menjadi tertarik dengan kinerja karyawan ini. Terutama untuk memastikan orang yang tepat dialokasikan untuk pekerjaan yang nyaman pada waktu yang

		<p>tepat. Dari sini, minat peran penambangan data (DM) telah berkembang yang tujuannya adalah penemuan pengetahuan dari sejumlah besar data. Dalam makalah ini, teknik DM digunakan untuk membangun model klasifikasi untuk memprediksi kinerja karyawan menggunakan dataset nyata yang dikumpulkan dari Kementerian Penerbangan Sipil Mesir (MOCA) melalui kuesioner yang disiapkan dan didistribusikan untuk 145 karyawan. Tiga teknik DM utama digunakan untuk membangun model klasifikasi dan mengidentifikasi faktor-faktor paling efektif yang secara positif mempengaruhi kinerja. Teknik-tekniknya adalah Decision Tree (DT), Naïve Bayes, dan Support Vector Machine (SVM). Untuk mendapatkan model yang sangat akurat, beberapa percobaan dilakukan berdasarkan teknik sebelumnya yang diimplementasikan dalam alat WEKA untuk memungkinkan pembuat keputusan dan profesional sumber daya manusia untuk memprediksi dan meningkatkan kinerja karyawan mereka.</p>
--	--	---

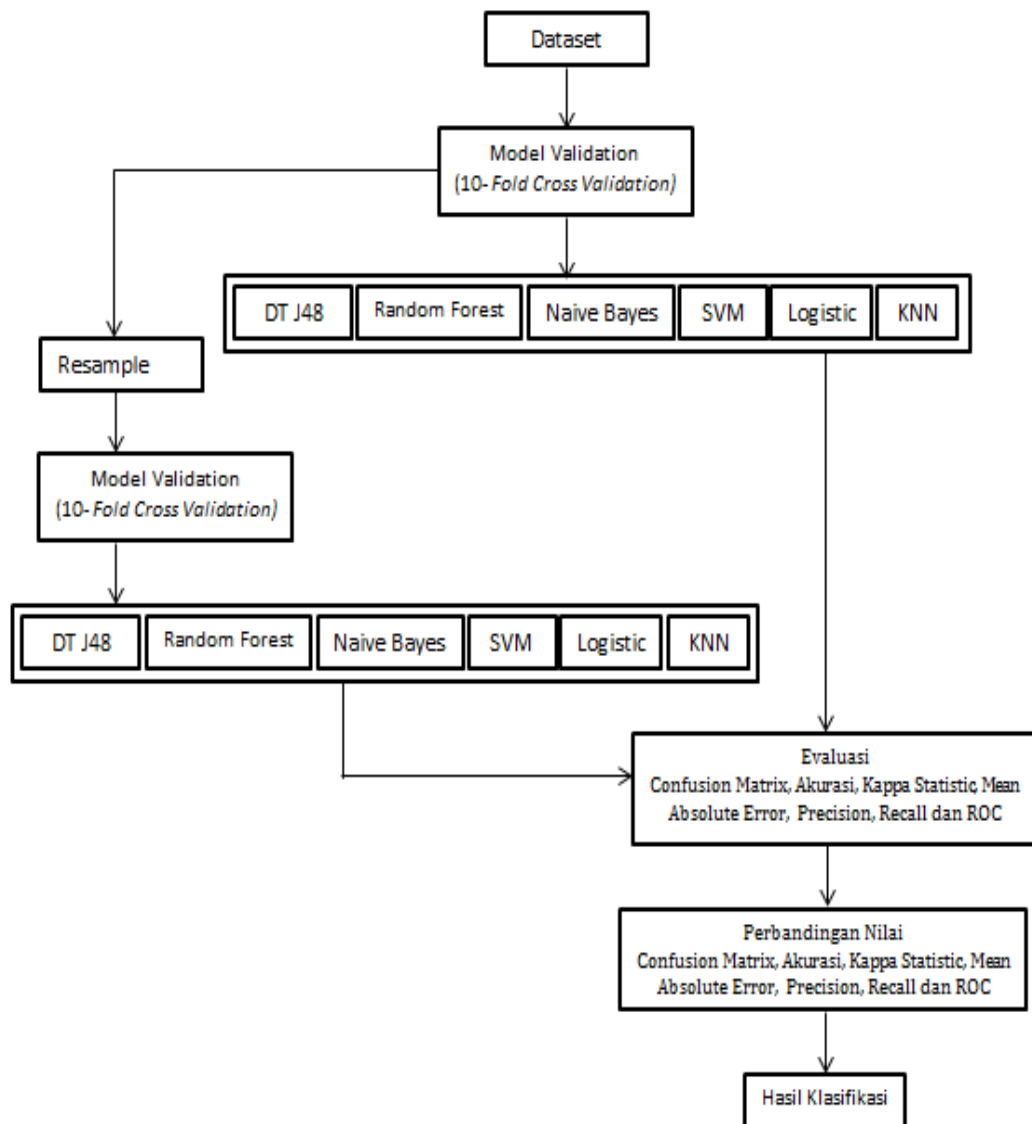
2	<p>Penerapan Resampling Dan Adaboost Untuk Penanganan Masalah Ketidakseimbangan Kelas Berbasis Naïve Bayes Pada Prediksi Churn Pelanggan[15]</p>	<p>Banyaknya operator seluler mendorong persaingan usaha yang sangat ketat. Kemudahan pelanggan untuk berpindah ke pesaing merupakan perhatian utama bagi bagian CRM (Customer Relationship Management), karena untuk mendapatkan pelanggan baru membutuhkan biaya yang jauh lebih mahal daripada mempertahankan pelanggan yang sudah ada. Untuk mengambil tindakan yang tepat dalam mempertahankan pelanggan harus mengetahui kecenderungan pelanggan apakah akan mengalami churn atau tidak. Prediksi kecenderungan pelanggan dilakukan dengan menggunakan model data mining. Pada penelitian ini akan diterapkan teknik resampling dan teknik ensemble AdaBoost untuk memperbaiki kinerja pengklasifikasi sedangkan untuk mengukur kinerja model digunakan software RapidMiner. Hasil penelitian menunjukkan bahwa model integrasi random oversampling, AdaBoost, dan Naïve Bayes memiliki kinerja yang lebih baik karena memiliki nilai AUC (Area Under the ROC (Receiver Operating Characteristic) Curve) yang lebih baik.</p>
---	--	---

## BAB III

### METODOLOGI PENELITIAN

#### 3.1. Metodologi Penelitian

Penelitian ini dilakukan melalui beberapa tahapan mulai dari *pra-proses* data, klasifikasi, validasi dan akurasi. Berikut kerangka pemikiran pada gambar III.1.



Gambar III.1 Metodologi Penelitian

Berdasarkan gambar III.1 langkah-langkah pemikiran di atas, dapat dijabarkan sebagai berikut :

1. *Dataset* terlebih dahulu di diuji dengan beberapa model algoritma klasifikasi yaitu *Decision Tree J48*, *Random Forest*, *Naive Bayes*, *Support Vector Machine (SVM)*, *Logistic Regression*, *K- Nearest Neighbord (KNN)* menggunakan *10-fold cross validation*.
2. *Dataset* diuji kembali menggunakan teknik *praprocessing* data yaitu *resample* dan diuji kembali menggunakan algoritma klasifikasi yaitu *Decision Tree J48*, *Random Forest*, *Naive Bayes*, *Support Vector Machine (SVM)*, *Logistic Regression*, *K- Nearest Neighbord (KNN)* menggunakan *10-fold cross validation*.
3. Berdasarkan evaluasi pengujian akan diperoleh nilai akurasi, dimana evaluasi yang digunakan menggunakan enam *measurement* yaitu berdasarkan evaluasi *Confusion matrix*, *Akurasi*, *Kappa Statistic*, *Mean Absolute Error*, *Precision*, *Recall* dan *ROC*.
4. Langkah selanjutnya yaitu perbandingan algoritma klasifikasi yaitu *Decision Tree J48*, *Random Forest*, *Naive Bayes*, *Support Vector Machine (SVM)*, *Logistic Regression*, *K- Nearest Neighbord (KNN)* dengan algoritma klasifikasi yaitu *Decision Tree J48*, *Random Forest*, *Naive Bayes*, *Support Vector Machine (SVM)*, *Logistic Regression*, *K- Nearest Neighbord (KNN)* dengan Teknik *resample*.
5. Setelah melakukan perbandingan algoritma maka didapat hasil pengujian *Confusion matrix*, *Akurasi*, *Kappa Statistic*, *Mean Absolute Error*, *Precision*, *Recall* dan *ROC*.

### 3.1.1 Jenis Penelitian

Metode penelitian adalah metode yang digunakan untuk melakukan pengumpulan data dalam suatu penelitian. Diantaranya yaitu survei, observasi, eksperimen, *case studies* dan lain-lain. Penelitian ini bertujuan untuk melakukan perbandingan dan evaluasi dengan algoritma klasifikasi menggunakan teknik *resample* dan algoritma klasifikasi tanpa menggunakan teknik *resample*.

### 3.1.2 Metode Pengumpulan Data

Metode pengumpulan data untuk mendapatkan data yang akan kita uji pada penelitian yaitu menggunakan data sekunder. Data sekunder adalah data yang dikumpulkan dan dianalisis oleh orang lain baik yang telah dipublikasikan maupun yang belum dipublikasikan. Jenis data yang digunakan adalah data sekunder dengan mengunduh data dari website medium.com HRM dataset.

Pada penelitian ini penulis menggunakan metode eksperimen yaitu penelitian yang melibatkan penyelidikan beberapa variable menggunakan tes tertentu yang dikendalikan sendiri oleh penulis untuk dapat melakukan pengklasifikasian dengan langkah-langkah sebagai berikut :

#### 1. Pengumpulan Data

Pengumpulan data ini merupakan data sekunder yang artinya data yang dihasilkan bukan dari orang pertama atau data yang bukan diusahakan sendiri. Dataset ini diambil dari link website <https://medium.com/@nafeea3000/human-resource-analytics-using-machine-learning-6a32392f6ec1>. Dimana memiliki 28 atribut dan 1200 data.

#### 2. Pengolahan Awal Data

Pengolahan awal data merupakan tahapan untuk permbersihan data, *preprocessing* merumuskan standar antropometri serta penyeimbang agar data yang didapat akurasi maksimal. Dengan melakukan normalisasi, *replace missing values*, konversi data.

#### 3. Metode yang diusulkan

Setelah tahapan *preprocessing*. Penulis mengusulkan metode klasifikasi yang akan dibuatkan sesuai dengan karakteristik dari dataset. Data yang akan diuji menggunakan *cross validation 10 folds*, data dibagi menjadi dua data pelatihan (*training dataset*) dan data pengujian (*testing dataset*).

#### 4. Eksperimen dan pengujian model

Langkah selanjutnya adalah melakukan eksperimen yang meliputi cara pemilihan arsitektur yang tepat dari model atau dari model yang diusulkan sehingga menghasilkan metode yang tepat.

## 5. Evaluasi dan Validasi Hasil

Setelah melakukan eksperimen dengan model yang terbaik, bagian selanjutnya adalah evaluasi dan validasi hasil dari model yang dibuat untuk mendapatkan informasi model yang akurat. Evaluasi dan validasi menggunakan metode *confusion matrix*.

File Edit View													
datanew.csv													
Relation: datanew													
No.	1: EmpNumber	2: Age	3: Gender	4: EducationBackground	5: MaritalStatus	6: EmpDepartment	7: EmpJobRole	8: BusinessTravelFrequency	9: DistanceFromHome	10: EmpEducationLevel	11: EmpEnvironmentSatisfaction	12: EmpHourlyF	
	Nominal	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	
1	E1001000	32.0	Male	Marketing	Single	Sales	Sales Execut...	Travel_Rarely	10.0	3.0	4.0	5	
2	E1001006	47.0	Male	Marketing	Single	Sales	Sales Execut...	Travel_Rarely	14.0	4.0	4.0	4	
3	E1001007	40.0	Male	Life Sciences	Married	Sales	Sales Execut...	Travel_Frequently	5.0	4.0	4.0	4	
4	E1001009	41.0	Male	Human Resources	Divorced	Human Resour...	Manager	Travel_Rarely	10.0	4.0	2.0	7	
5	E1001010	60.0	Male	Marketing	Single	Sales	Sales Execut...	Travel_Rarely	16.0	4.0	1.0	8	
6	E1001011	27.0	Male	Life Sciences	Divorced	Development	Developer	Travel_Frequently	10.0	2.0	4.0	3	
7	E1001016	50.0	Male	Marketing	Married	Sales	Sales Repre...	Travel_Rarely	8.0	4.0	4.0	5	
8	E1001019	28.0	Female	Life Sciences	Single	Development	Developer	Travel_Rarely	1.0	2.0	1.0	6	
9	E1001020	36.0	Female	Life Sciences	Married	Development	Developer	Non-Travel	8.0	3.0	1.0	6	
10	E1001021	38.0	Female	Life Sciences	Single	Development	Developer	Travel_Rarely	1.0	3.0	3.0	8	
11	E1001022	44.0	Male	Medical	Single	Development	Developer	Non-Travel	24.0	3.0	1.0	4	
12	E1001024	47.0	Female	Medical	Divorced	Sales	Sales Execut...	Travel_Frequently	3.0	3.0	4.0	4	
13	E1001025	30.0	Male	Marketing	Divorced	Sales	Sales Execut...	Travel_Rarely	27.0	5.0	3.0	9	
14	E1001027	29.0	Male	Life Sciences	Single	Sales	Sales Repre...	Travel_Rarely	10.0	3.0	3.0	9	
15	E1001030	42.0	Male	Medical	Divorced	Development	Developer	Travel_Frequently	19.0	3.0	3.0	5	
16	E1001035	34.0	Female	Medical	Single	Development	Developer	Travel_Rarely	8.0	2.0	2.0	9	
17	E1001038	39.0	Female	Human Resources	Married	Human Resour...	Human Res...	Travel_Rarely	3.0	3.0	3.0	4	
18	E1001040	56.0	Male	Medical	Married	Development	Developer	Travel_Rarely	9.0	3.0	3.0	8	
19	E1001041	40.0	Female	Medical	Single	Development	Developer	Travel_Rarely	2.0	1.0	4.0	8	
20	E1001042	27.0	Female	Medical	Single	Development	Developer	Travel_Rarely	7.0	3.0	4.0	5	
21	E1001044	29.0	Male	Marketing	Divorced	Sales	Sales Repre...	Travel_Rarely	10.0	3.0	4.0	8	
22	E1001047	53.0	Male	Life Sciences	Single	Development	Developer	Travel_Rarely	6.0	3.0	4.0	8	
23	E1001049	35.0	Female	Life Sciences	Divorced	Development	Senior Devel...	Non-Travel	2.0	4.0	4.0	6	
24	E1001050	32.0	Male	Life Sciences	Married	Development	Developer	Travel_Frequently	24.0	4.0	1.0	8	
25	E1001053	34.0	Female	Life Sciences	Divorced	Development	Developer	Travel_Rarely	8.0	5.0	2.0	3	
26	E1001054	52.0	Male	Marketing	Married	Sales	Manager	Travel_Rarely	3.0	4.0	3.0	3	
27	E1001058	33.0	Male	Other	Single	Development	Developer	Travel_Rarely	1.0	4.0	4.0	6	
28	E1001059	25.0	Female	Medical	Single	Sales	Sales Execut...	Travel_Rarely	26.0	1.0	3.0	3	
29	E1001061	45.0	Male	Technical Degree	Single	Sales	Sales Repre...	Travel_Rarely	2.0	2.0	2.0	4	
30	E1001062	23.0	Male	Medical	Single	Development	Developer	Travel_Rarely	10.0	1.0	1.0	7	

Gambar III.2. Sampel Dataset



File Edit View											
datanew.csv											
Relation: datanew											
14: EmpJobInvolvement 15: EmpJobLevel 16: EmpJobSatisfaction 17: OverTime 18: EmpLastSalaryHikePercent 19: EmpRelationshipSatisfaction 20: TotalWorkExperienceInYears 21: TrainingTimesLastYear											
Numeric	Numeric	Numeric	Numeric	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
3.0	2.0	4.0	1.0	No	12.0	4.0	10.0	2.0	2.0	2.0	2.0
3.0	2.0	1.0	2.0	No	12.0	4.0	20.0	2.0	2.0	2.0	2.0
2.0	3.0	1.0	5.0	Yes	21.0	3.0	20.0	2.0	2.0	2.0	2.0
2.0	5.0	4.0	3.0	No	15.0	2.0	23.0	2.0	2.0	2.0	2.0
3.0	2.0	1.0	8.0	No	14.0	4.0	10.0	1.0	1.0	1.0	1.0
3.0	3.0	1.0	1.0	No	21.0	3.0	9.0	4.0	4.0	4.0	4.0
3.0	1.0	2.0	7.0	No	15.0	4.0	4.0	2.0	2.0	2.0	2.0
1.0	1.0	2.0	7.0	Yes	13.0	4.0	10.0	4.0	4.0	4.0	4.0
4.0	3.0	1.0	9.0	No	14.0	1.0	10.0	2.0	2.0	2.0	2.0
3.0	3.0	3.0	4.0	Yes	14.0	4.0	10.0	4.0	4.0	4.0	4.0
1.0	1.0	3.0	2.0	No	14.0	3.0	9.0	5.0	5.0	5.0	5.0
3.0	4.0	3.0	9.0	Yes	12.0	4.0	28.0	2.0	2.0	2.0	2.0
3.0	2.0	4.0	7.0	No	23.0	4.0	10.0	2.0	2.0	2.0	2.0
3.0	1.0	3.0	1.0	No	11.0	3.0	1.0	6.0	6.0	6.0	6.0
4.0	1.0	3.0	6.0	Yes	12.0	4.0	7.0	2.0	2.0	2.0	2.0
3.0	2.0	3.0	3.0	No	11.0	4.0	10.0	2.0	2.0	2.0	2.0
4.0	2.0	2.0	9.0	No	15.0	3.0	12.0	3.0	3.0	3.0	3.0
3.0	4.0	4.0	7.0	No	11.0	3.0	30.0	1.0	1.0	1.0	1.0
2.0	1.0	4.0	0.0	Yes	20.0	4.0	5.0	4.0	4.0	4.0	4.0
2.0	2.0	1.0	8.0	No	19.0	1.0	9.0	2.0	2.0	2.0	2.0
3.0	1.0	2.0	1.0	No	14.0	4.0	2.0	2.0	2.0	2.0	2.0
3.0	2.0	4.0	2.0	No	17.0	4.0	19.0	4.0	4.0	4.0	4.0
3.0	2.0	1.0	2.0	No	11.0	1.0	16.0	2.0	2.0	2.0	2.0
3.0	2.0	4.0	1.0	No	15.0	4.0	10.0	2.0	2.0	2.0	2.0
3.0	2.0	1.0	3.0	No	14.0	3.0	7.0	3.0	3.0	3.0	3.0
2.0	4.0	1.0	1.0	No	11.0	1.0	34.0	3.0	3.0	3.0	3.0
3.0	1.0	4.0	1.0	Yes	13.0	3.0	10.0	2.0	2.0	2.0	2.0
3.0	2.0	4.0	1.0	No	23.0	2.0	6.0	5.0	5.0	5.0	5.0
1.0	2.0	3.0	0.0	No	13.0	1.0	9.0	3.0	3.0	3.0	3.0
4.0	1.0	3.0	1.0	No	18.0	4.0	2.0	3.0	3.0	3.0	3.0
4.0	2.0	3.0	4.0	Yes	12.0	4.0	8.0	2.0	2.0	2.0	2.0
3.0	2.0	4.0	1.0	No	22.0	3.0	6.0	5.0	5.0	5.0	5.0
2.0	3.0	1.0	1.0	No	15.0	2.0	10.0	1.0	1.0	1.0	1.0
3.0	1.0	4.0	0.0	No	18.0	1.0	5.0	3.0	3.0	3.0	3.0
1.0	1.0	4.0	3.0	No	14.0	1.0	5.0	2.0	2.0	2.0	2.0
3.0	2.0	3.0	4.0	No	12.0	1.0	11.0	3.0	3.0	3.0	3.0
2.0	2.0	1.0	6.0	No	17.0	3.0	6.0	2.0	2.0	2.0	2.0

Gambar III.3 Sample Dataset Lanjutan 1

File Edit View											
datanew.csv											
Relation: datanew											
21: TrainingTimesLastYear 22: EmpWorkLifeBalance 23: ExperienceYearsAtThisCompany 24: ExperienceYearsInCurrentRole 25: YearsSinceLastPromotion 26: YearsWithCurrentManager 27: Attrition 28: PerformanceRating											
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal	Nominal	Nominal
10.0	2.0	2.0	10.0	7.0	0.0	0.0	8.0	No	Luar Bias	Luar Bias	Luar Bias
20.0	2.0	3.0	7.0	7.0	1.0	7.0	7.0	No	Luar Bias	Luar Bias	Luar Bias
20.0	2.0	3.0	18.0	13.0	1.0	12.0	12.0	No	Sangat Luar Biasa	Sangat Luar Biasa	Sangat Luar Biasa
23.0	2.0	2.0	21.0	6.0	12.0	6.0	6.0	No	Luar Bias	Luar Bias	Luar Bias
10.0	1.0	3.0	2.0	2.0	2.0	2.0	2.0	No	Luar Bias	Luar Bias	Luar Bias
9.0	4.0	2.0	9.0	7.0	1.0	7.0	7.0	No	Sangat Luar Biasa	Sangat Luar Biasa	Sangat Luar Biasa
4.0	2.0	3.0	2.0	2.0	2.0	2.0	2.0	No	Luar Bias	Luar Bias	Luar Bias
10.0	4.0	3.0	7.0	7.0	3.0	7.0	7.0	Yes	Luar Bias	Luar Bias	Luar Bias
10.0	2.0	3.0	8.0	7.0	0.0	5.0	5.0	No	Luar Bias	Luar Bias	Luar Bias
10.0	4.0	4.0	1.0	0.0	0.0	0.0	0.0	No	Luar Bias	Luar Bias	Luar Bias
9.0	5.0	3.0	5.0	2.0	1.0	4.0	4.0	No	Luar Bias	Luar Bias	Luar Bias
28.0	2.0	2.0	22.0	2.0	11.0	12.0	12.0	No	Luar Bias	Luar Bias	Luar Bias
10.0	2.0	2.0	8.0	7.0	7.0	7.0	7.0	No	Sangat Luar Biasa	Sangat Luar Biasa	Sangat Luar Biasa
1.0	6.0	3.0	1.0	0.0	0.0	0.0	0.0	No	Luar Bias	Luar Bias	Luar Bias
7.0	2.0	3.0	2.0	2.0	2.0	2.0	2.0	Yes	Luar Bias	Luar Bias	Luar Bias
10.0	2.0	3.0	5.0	1.0	4.0	3.0	3.0	No	Luar Bias	Luar Bias	Luar Bias
12.0	3.0	1.0	8.0	3.0	3.0	6.0	6.0	No	Luar Bias	Luar Bias	Luar Bias
30.0	1.0	2.0	10.0	7.0	1.0	1.0	1.0	No	Luar Bias	Luar Bias	Luar Bias
5.0	2.0	2.0	4.0	2.0	2.0	3.0	3.0	No	Sangat Luar Biasa	Sangat Luar Biasa	Sangat Luar Biasa
9.0	2.0	1.0	7.0	6.0	0.0	7.0	7.0	No	Luar Bias	Luar Bias	Luar Bias
2.0	2.0	3.0	2.0	2.0	2.0	2.0	2.0	No	Luar Bias	Luar Bias	Luar Bias
19.0	4.0	3.0	2.0	2.0	2.0	2.0	2.0	No	Luar Bias	Luar Bias	Luar Bias
16.0	2.0	4.0	1.0	0.0	0.0	0.0	0.0	No	Luar Bias	Luar Bias	Luar Bias
10.0	2.0	3.0	10.0	8.0	4.0	7.0	7.0	No	Luar Bias	Luar Bias	Luar Bias
7.0	3.0	3.0	0.0	0.0	0.0	0.0	0.0	No	Luar Bias	Luar Bias	Luar Bias
24.0	3.0	4.0	24.0	6.0	1.0	16.0	16.0	No	Sangat Luar Biasa	Sangat Luar Biasa	Sangat Luar Biasa
10.0	2.0	2.0	10.0	9.0	7.0	8.0	8.0	Yes	Luar Bias	Luar Bias	Luar Bias
6.0	5.0	2.0	6.0	5.0	1.0	4.0	4.0	No	Sangat Luar Biasa	Sangat Luar Biasa	Sangat Luar Biasa
9.0	3.0	3.0	8.0	7.0	3.0	1.0	1.0	No	Luar Bias	Luar Bias	Luar Bias
2.0	3.0	3.0	2.0	2.0	0.0	2.0	2.0	No	Luar Bias	Luar Bias	Luar Bias
8.0	2.0	3.0	5.0	4.0	1.0	3.0	3.0	Yes	Luar Bias	Luar Bias	Luar Bias
6.0	5.0	3.0	6.0	5.0	1.0	4.0	4.0	No	Sangat Luar Biasa	Sangat Luar Biasa	Sangat Luar Biasa
10.0	1.0	3.0	10.0	7.0	0.0	9.0	9.0	No	Luar Bias	Luar Bias	Luar Bias
5.0	3.0	3.0	4.0	3.0	1.0	2.0	2.0	No	Luar Bias	Luar Bias	Luar Bias
5.0	2.0	3.0	2.0	2.0	2.0	0.0	0.0	No	Luar Bias	Luar Bias	Luar Bias
11.0	3.0	2.0	8.0	7.0	1.0	1.0	1.0	No	Luar Bias	Luar Bias	Luar Bias
--	--	--	--	--	--	--	--	--	--	--	--

Gambar III.4. Sample Dataset Lanjutan 2

## BAB IV

### HASIL DAN PEMBAHASAN

#### 4.1. Dataset

Penelitian ini menggunakan *dataset* HRM. Tahap preprocessing dilakukan dengan melihat beberapa ketentuan awal yang perlu diperhatikan diantaranya jumlah atribut, jumlah modul dan jumlah cacat pada setiap *dataset*. Spesifikasi *dataset* yang digunakan dalam penelitian ini sebagai berikut.

Tabel IV.1  
Spesifikasi *dataset*

<i>Dataset</i>	Jumlah Atribut	Jumlah Data
<b>Data HRM_431</b>	<b>27</b>	<b>1200</b>

Tabel IV.2  
Penjelasan Atribut pada *dataset* HRM

Atribut	Description
<i>Age</i>	Umur
<i>Gender</i>	Male, Female
<i>Education Background</i>	Latar Belakang Pendidikan
<i>Marital Status</i>	Status Pernikahan
<i>Emp Depertment</i>	-
<i>Emp Job Role</i>	Peran Pekerjaan
<i>Business Travel Frequency</i>	Frekuensi Perjalanan Bisnis
<i>Distance From Home</i>	Jarak Dari Rumah
<i>Emp Education Level</i>	Tingkatan Pendidikan
<i>Emp Environment Satisfaction</i>	Kepuasan Lingkungan
<i>Emp Hourly Rate</i>	Tarif per Jam

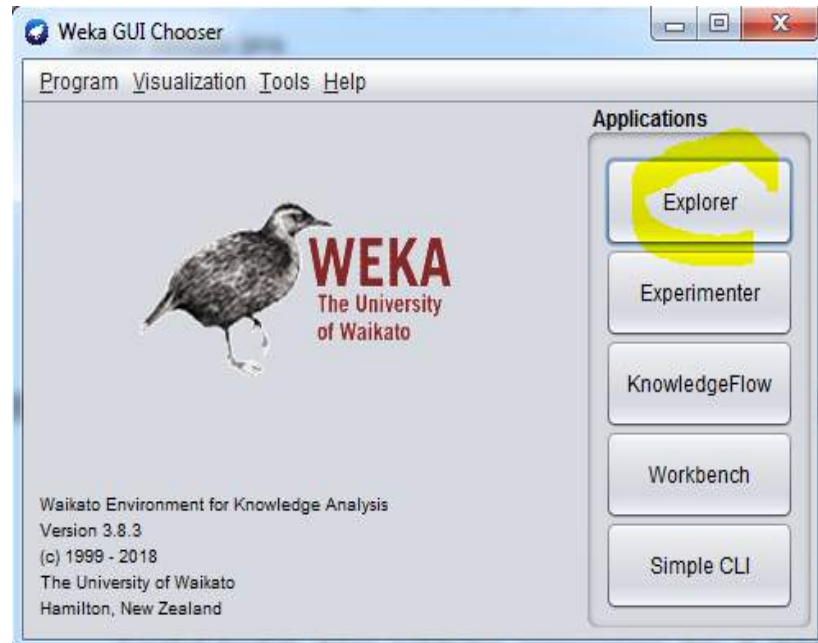
<i>Emp Job Involvement</i>	Keterlibatan Pekerjaan
<i>Emp Job Level</i>	Tingkat Pekerjaan
<i>Emp Job Satisfaction</i>	Kepuasan Kerja
<i>Num Companies Worked</i>	Banyak Perusahaan Bekerja
<i>Overtime</i>	Lembur
<i>Emp Salary Hike Percent</i>	Kenaikan Gaji Porsen
<i>Emp Relationship Satisfaction</i>	Kepuasan Hubungan
<i>Total Work Experience In Years</i>	Total Pengalaman Kerja Dalam Beberapa Tahun
<i>Training Time Last Year</i>	Waktu Pelatihan Tahun Terakhir
<i>Emp Work Life Balance</i>	Neraca Kehidupan Kerja
<i>Experience Years At This Company</i>	Pengalaman Bertahun-Tahun Di Perusahaan Ini
<i>Experience Years In Current Role</i>	Pengalaman Tahun Dalam Peran Saat Ini
<i>Years Since Last Promotion</i>	Tahun Sejak Promosi Terakhir
<i>Attrition</i>	Erosi
<i>Performance Rating</i>	Peringkat Kinerja

#### 4.2. Preprocessing Data

Tahap berikutnya yaitu melakukan pengujian dataset dengan beberapa algoritma klasifikasi diantaranya Desicion Tree J48, Random Forest, Naive Bayes, KNN, Logistic dan SVM. Kemudian hasil dari ke enam jenis algoritma yang diusulkan akan dievaluasi menggunakan Confusion Matrix, Akurasi, Kappa Statistic, Mean Absolute Error, Precision, Recall dan ROC. Tools yang digunakan dalam pengujian algoritma yaitu aplikasi WEKA (*Waikato Environment for Knowledge Analysis*) juga merupakan aplikasi data mining open source berbasis java.

Langkah-langkah dalam melakukan *preprocessing data* di WEKA sebagai berikut :

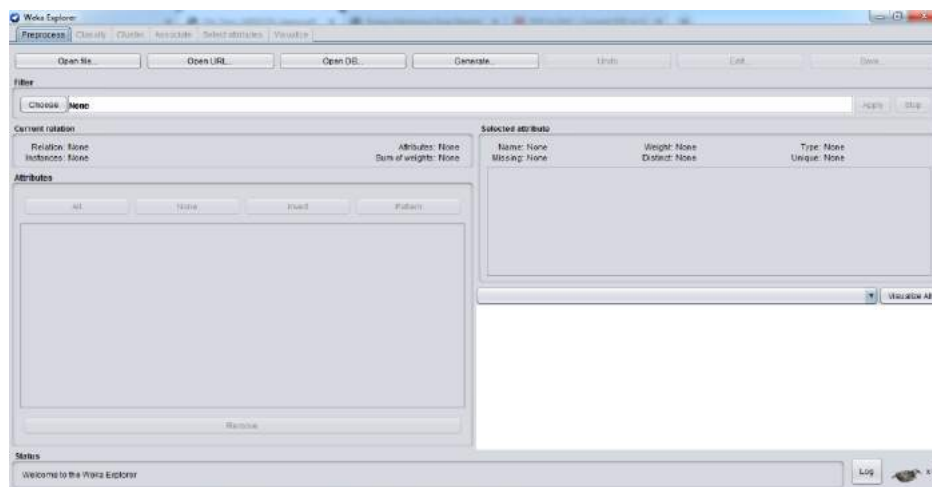
- Buka dan jalankan aplikasi WEKA



Gambar IV.1. Tampilan awal WEKA

Pada gambar 4.1 diatas merupakan tampilan awal dari aplikasi WEKA (*Waikato Environment for Knowledge Analysis*) 3.8.3 yang dijadikan sebagai tools untuk menguji dataset HRM menggunakan *Cross- Validation*.

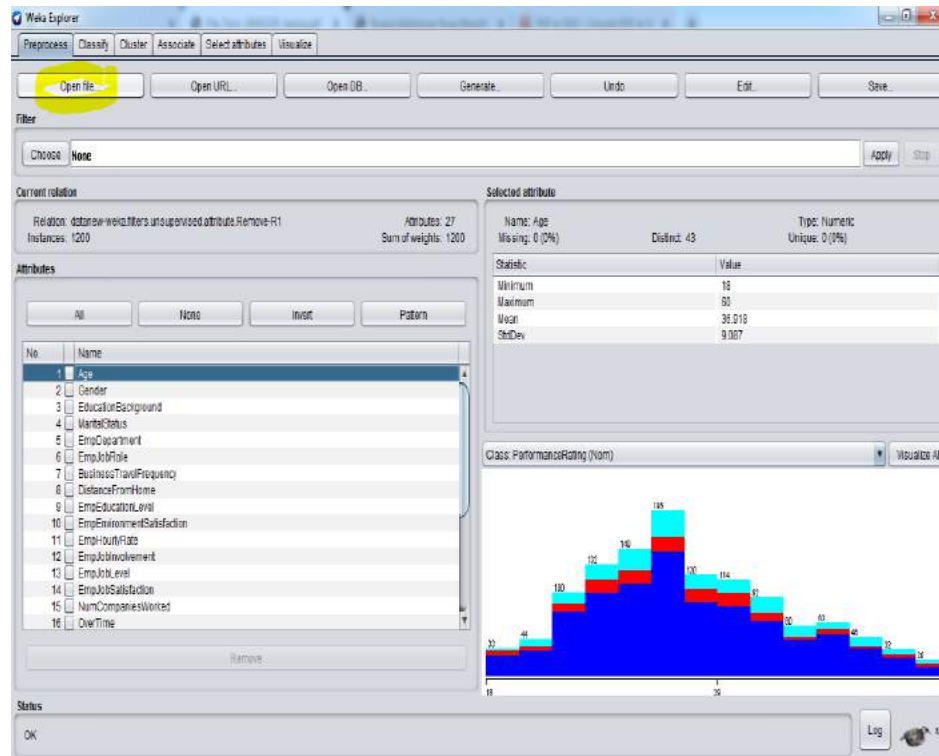
- Klik button Explorer untuk membuka antarmuka WEKA Explorer



Gambar IV.2. Tampilan antarmuka WEKA

Pada gambar 4.2. diatas merupakan tampilan antarmuka *preprocess* weka untuk memulai preprocessing dataset.

- Langkah selanjutnya buka dataset yang akan diuji

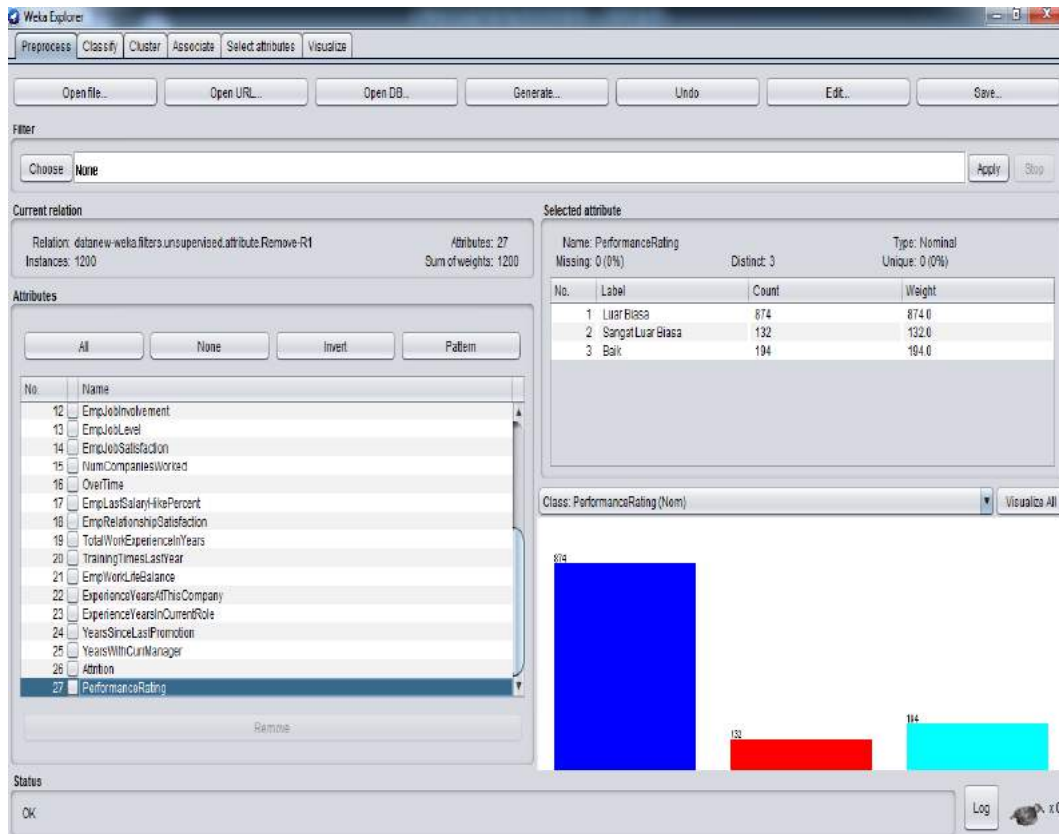


Gambar IV.3. Tampilan Dataset di WEKA

Pada gambar 4.3. diatas merupakan tampilan dataset yang akan diuji dengan cara mengklik button open file kemudian cari file dataset yang akan diuji, maka tampilan datasetnya akan muncul seperti gambar diatas.

#### 4.2.1. Preprocessing Data Sebelum di Resample

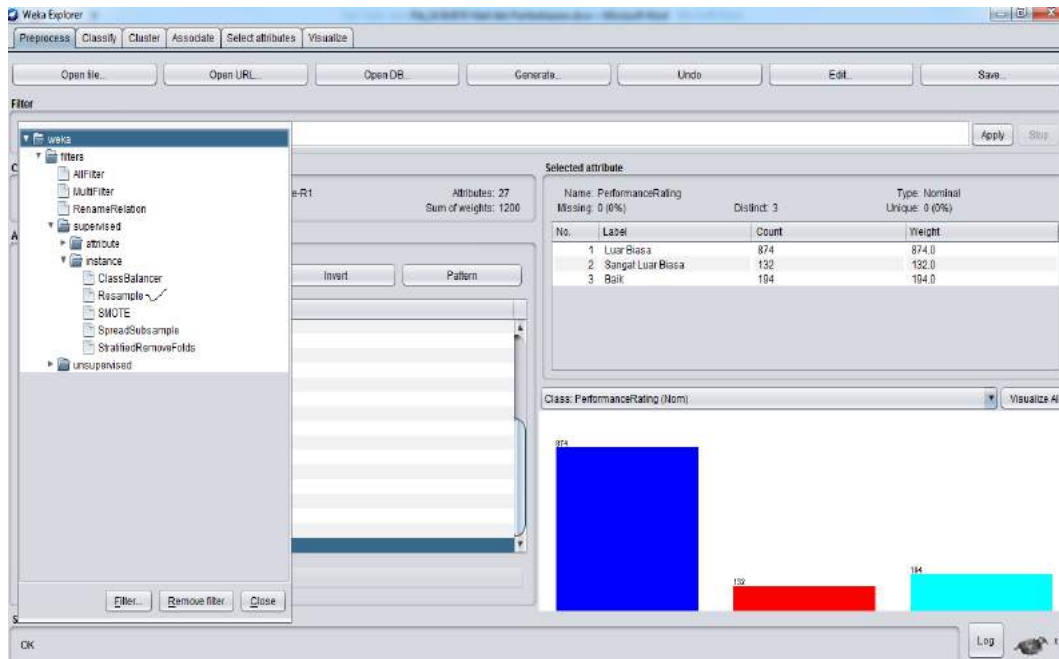
Tampilan awal saat proses preprocessing data sebelum di resample, terlihat pada gambar pada grafik batang dibawah bahwa label luar biasa memiliki *count* tertinggi dibanding dua label lainnya sehingga data tersebut harus dibalance kan terlebih dahulu.



Gambar IV.4. Tampilan Preprocessing Data Sebelum di Resample

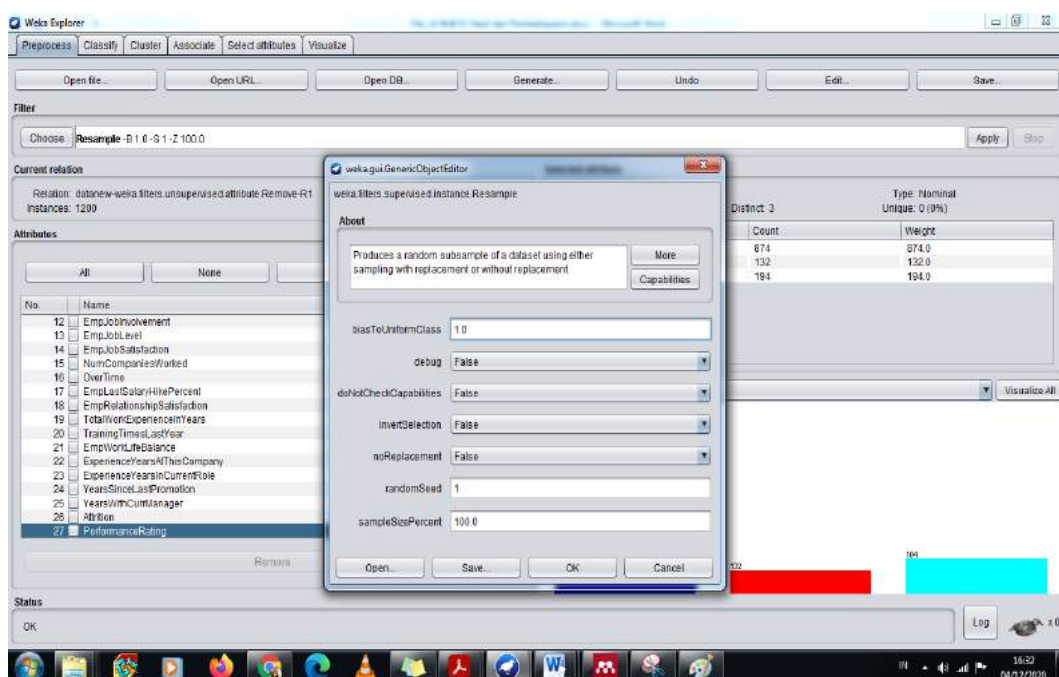
#### 4.2.2. Proses Preprocessing Data di Resample

Pada halaman antarmuka di weka langkah yang dilakukan setelah dataset yang akan diuji sudah dibuka selanjutnya adalah di menu filter kita pilih button choose untuk memilih teknik resample didalam folder instance kita klik.



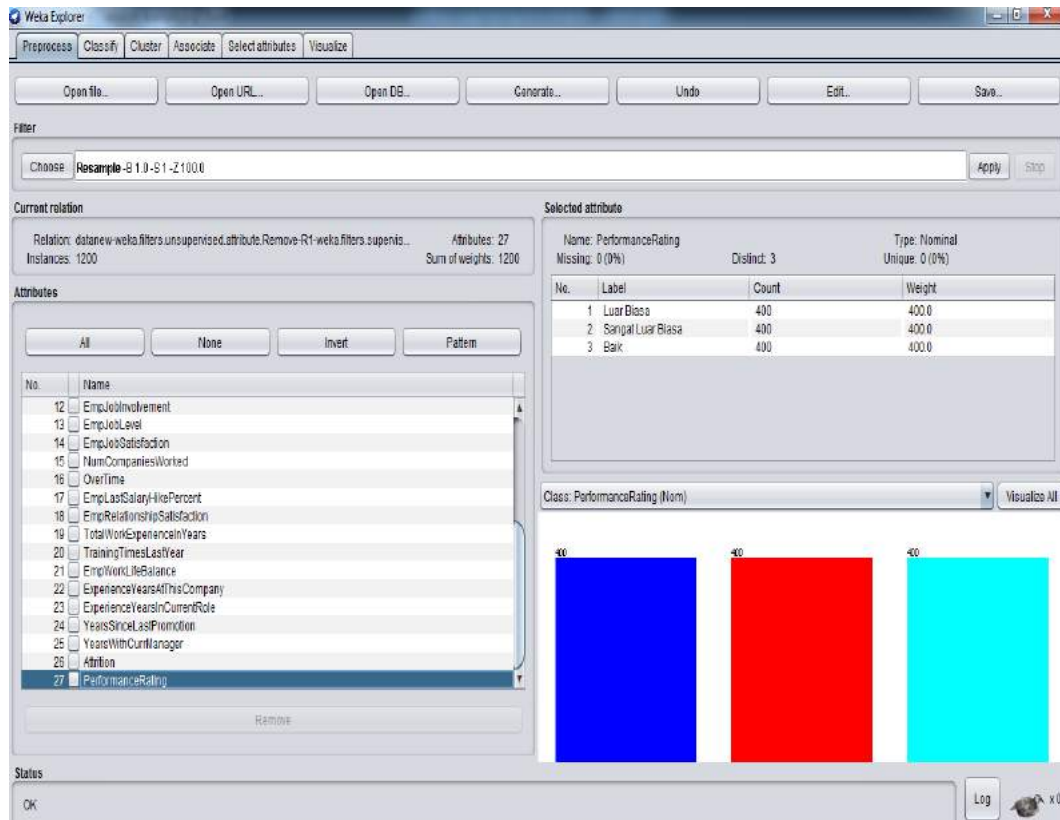
Gambar IV.5. Tampilan Proses Preprocessing Data di Resample

Kemudian *double* klik pada text box disamping *button choose* di menu filter akan muncul tampilan *weka.gui.GenericObjectEditor* lalu kita ubah di parameter *biasToUniformClass* yang semula 0.0 kita rubah menjadi 1.0 seperti gambar dibawah ini :

Gambar IV.6. Tampilan *weka.gui.GenericObjectEditor*

#### 4.2.3. Preprocessing Data Sesudah di Resample

Tampilan olah dataset ditahap preprocessing setelah dilakukan teknik resample dataset terlihat pada gambar grafik batang dibawah bahwa ketiga label tersebut sudah *balance* .

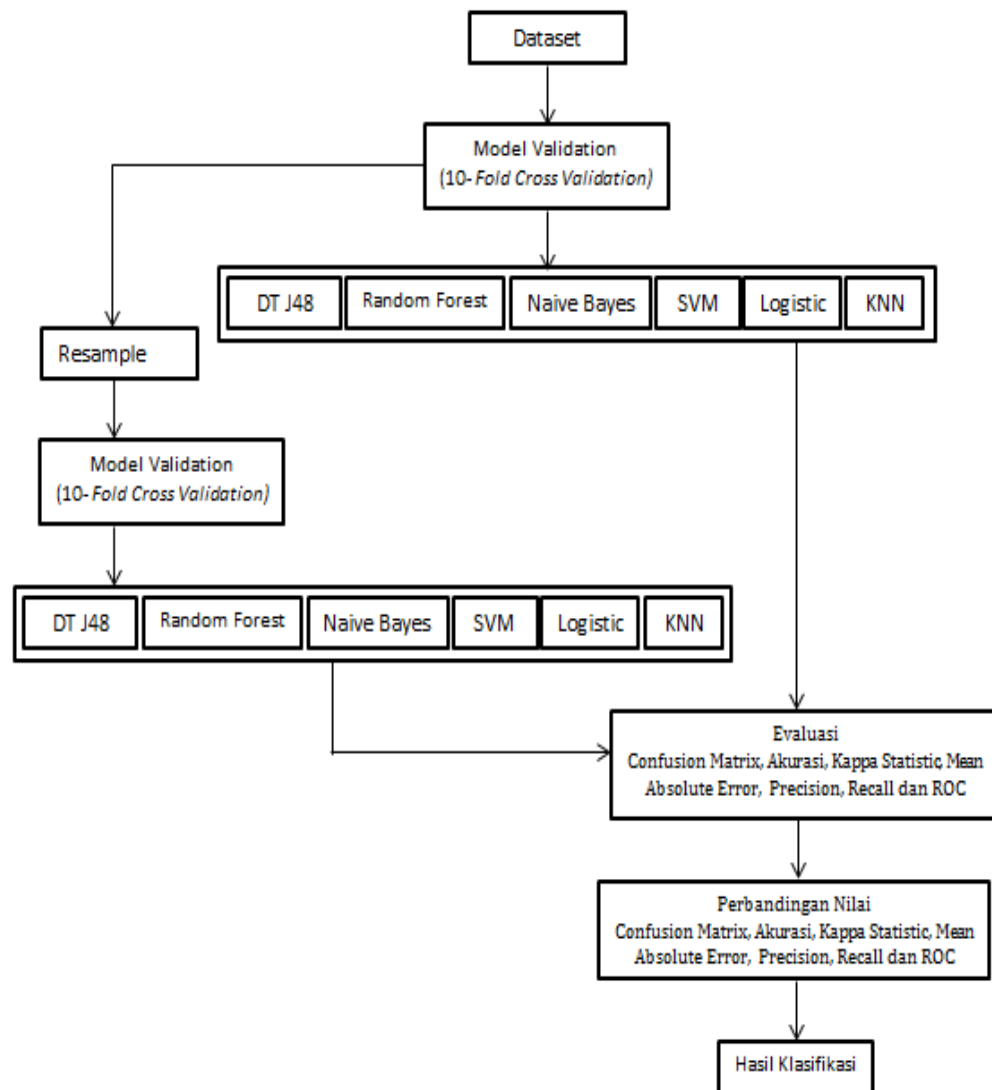


Gambar IV.7. Tampilan Preprocessing Data Sesudah di Resample

#### 4.3. Eksperimen dan Pengujian Model

Pada fase eksperimen dan pengujian model melibatkan teknik data mining yaitu dengan cara melakukan pemilihan teknik data mining dan menetapkan algoritma yang akan digunakan. *Tool* yang digunakan pada fase ini adalah WEKA versi 3.8.3. Adapun hasil dalam pengujian model yang akan dilakukan adalah mengklasifikasi Luar biasa, Sangat luar biasa, dan baik menggunakan algoritma *Desicion Tree J48*, *Random Forest*, *Naive Bayes*, *KNN*, *Logistic* dan *SVM*. Berikut langkah-langkah pengujian dan perbandingan grafik dataset pada tools Weka 3.8.3.





Gambar IV.8. Alur Metode yang diusulkan

#### 4.3.1. Pengujian Model Klasifikasi

Pada bagian ini penulis melakukan percobaan pada dataset HRM dengan menggunakan model klasifikasi.

1. Hasil *Confusion Matrix* Algoritma *Decision Tree J48*

Tabel IV.3

Hasil *Confusion Matrix* Algoritma *Decision Tree J48*

<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
854	3	17	<b>a = Luar Biasa</b>
18	109	5	<b>b = Sangat Luar Biasa</b>
11	1	182	<b>c = Baik</b>

2. Hasil *Confusion Matrix* Algoritma *Random Forest*

Tabel IV.4

Hasil *Confusion Matrix* Algoritma *Random Forest*

<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
861	1	12	<b>a = Luar Biasa</b>
27	104	1	<b>b = Sangat Luar Biasa</b>
19	0	175	<b>c = Baik</b>

3. Hasil *Confusion Matrix* Algoritma *Naive Bayes*

Tabel IV.5

Hasil *Confusion Matrix* Algoritma *Naive Bayes*

<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
786	28	60	<b>a = Luar Biasa</b>
46	81	5	<b>b = Sangat Luar Biasa</b>
97	2	95	<b>c = Baik</b>

4. Hasil *Confusion Matrix* Algoritma *Support Vector Machine (SVM)*

Tabel IV.6

Hasil *Confusion Matrix* Algoritma *Support Vector Machine (SVM)*

<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
792	13	64	<b>a = Luar Biasa</b>
39	91	2	<b>b = Sangat Luar Biasa</b>
91	0	103	<b>c = Baik</b>

5. Hasil *Confusion Matrix* Algoritma *Logistic*

Tabel IV.7

Hasil *Confusion Matrix* Algoritma *Logistic*

<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
786	24	64	<b>a = Luar Biasa</b>
35	93	4	<b>b = Sangat Luar Biasa</b>
87	5	102	<b>c = Baik</b>

6. Hasil *Confusion Matrix* Algoritma *K-Nearest Neighbor*

Tabel IV.8

Hasil *Confusion Matrix* Algoritma *K-Nearest Neighbor*

<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
809	27	38	<b>a = Luar Biasa</b>
37	91	4	<b>b = Sangat Luar Biasa</b>
43	7	144	<b>c = Baik</b>

Tabel IV.9.

Hasil Kinerja Klasifikasi dari algoritma *Decision Tree J48*, *Random Forest*, *SVM*,  
*Naive Bayes*, *Logistic*, *KNN*

Algoritma	Akurasi	Kappa Statistic	MAE	Precision	Recall	ROC
<b><i>DT J48</i></b>	<b>92,83%</b>	<b>0,8316</b>	<b>0,0669</b>	<b>0,928</b>	<b>0,928</b>	<b>0,928</b>
<i>Random Forest</i>	90,41%	0,759	0,1669	0,906	0,904	0,944
<i>Naive Bayes</i>	79,16%	0,4776	0,1685	0,783	0,792	0,862
<i>SVM</i>	82, 33%	0,5562	0,2659	0,817	0,823	0,781
<i>Logistic</i>	81,25%	0,5522	0,1588	0,809	0,813	0,908
<i>KNN</i>	66,91%	0,1755	0,2212	0,646	0,669	0,580

Hasil kinerja klasifikasi pada tabel diatas menunjukkan hasil akurasi tertinggi yaitu algoritma *Decision Tree J48* dengan nilai akurasi sebesar 92,83%.

#### 4.3.2. Pengujian Model Klasifikasi dengan Resample

Pada bagian ini penulis melakukan percobaan pada dataset HRM dengan menggunakan penggabungan model yaitu model klasifikasi dan Teknik *Resample*.

##### 1. Hasil *Confusion Matrix* Algoritma *Decision Tree J48* dengan *Resample*

Tabel IV.10

Hasil *Confusion Matrix* Algoritma *Decision Tree J48* dengan *Resample*

a	b	c	Classified as
854	3	17	<b>a = Luar Biasa</b>
18	109	5	<b>b = Sangat Luar Biasa</b>
11	1	182	<b>c = Baik</b>

2. Hasil *Confusion Matrix* Algoritma *Random Forest* dengan *Resample*

Tabel IV.11

Hasil *Confusion Matrix* Algoritma *Random Forest* dengan *Resample*

<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
861	1	12	<b>a = Luar Biasa</b>
27	104	1	<b>b = Sangat Luar Biasa</b>
19	0	175	<b>c = Baik</b>

3. Hasil *Confusion Matrix* Algoritma *Naive Bayes* dengan *Resample*

Tabel IV.12

Hasil *Confusion Matrix* Algoritma *Naive Bayes* dengan *Resample*

<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
786	28	60	<b>a = Luar Biasa</b>
46	81	5	<b>b = Sangat Luar Biasa</b>
97	2	95	<b>c = Baik</b>

4. Hasil *Confusion Matrix* Algoritma *Support Vector Machine (SVM)* dengan *Resample*

Tabel IV.13

Hasil *Confusion Matrix* Algoritma *Support Vector Machine (SVM)* dengan *Resample*

<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
792	13	64	<b>a = Luar Biasa</b>
39	91	2	<b>b = Sangat Luar Biasa</b>
91	0	103	<b>c = Baik</b>

5. Hasil *Confusion Matrix* Algoritma *Logistic* dengan *Resample*

Tabel IV.14

Hasil *Confusion Matrix* Algoritma *Logistic* dengan *Resample*

<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
786	24	64	<b>a = Luar Biasa</b>
35	93	4	<b>b = Sangat Luar Biasa</b>
87	5	102	<b>c = Baik</b>

6. Hasil *Confusion Matrix* Algoritma *K-Nearest Neighbor* dengan *Resample*

Tabel IV.15

Hasil *Confusion Matrix* Algoritma *K-Nearest Neighbor* dengan *Resample*

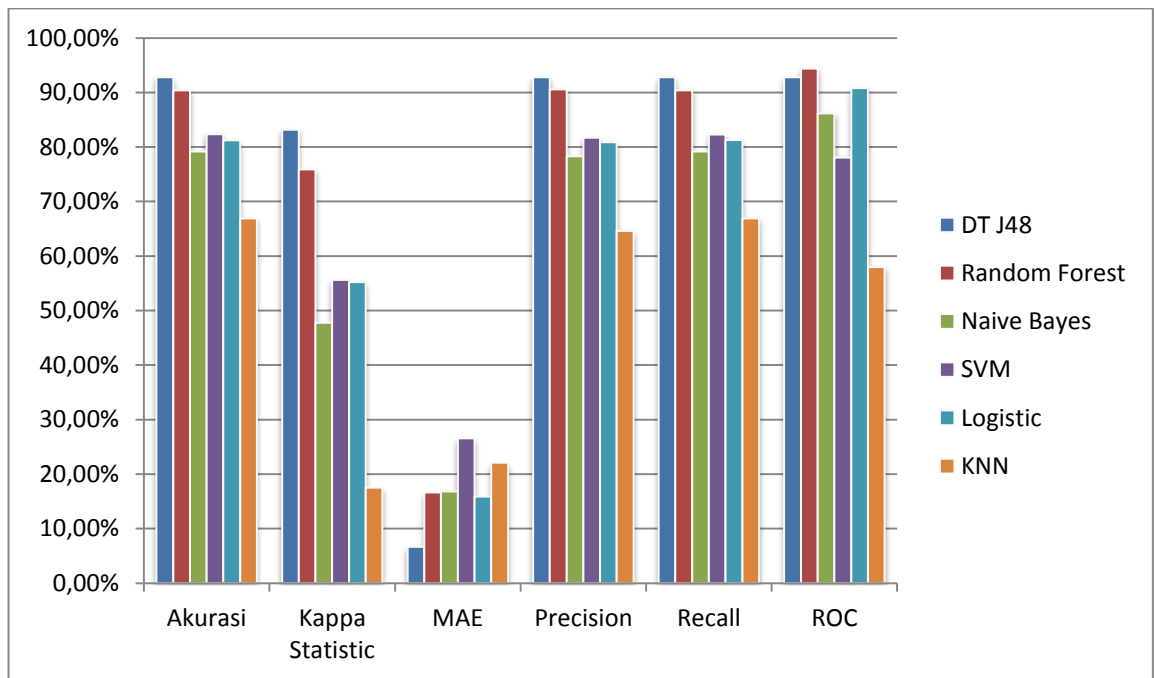
<b>a</b>	<b>b</b>	<b>c</b>	<b>Classified as</b>
809	27	38	<b>a = Luar Biasa</b>
37	91	4	<b>b = Sangat Luar Biasa</b>
43	7	144	<b>c = Baik</b>

Tabel IV.16.

Hasil Kinerja Pengujian Klasifikasi dengan *Resample*

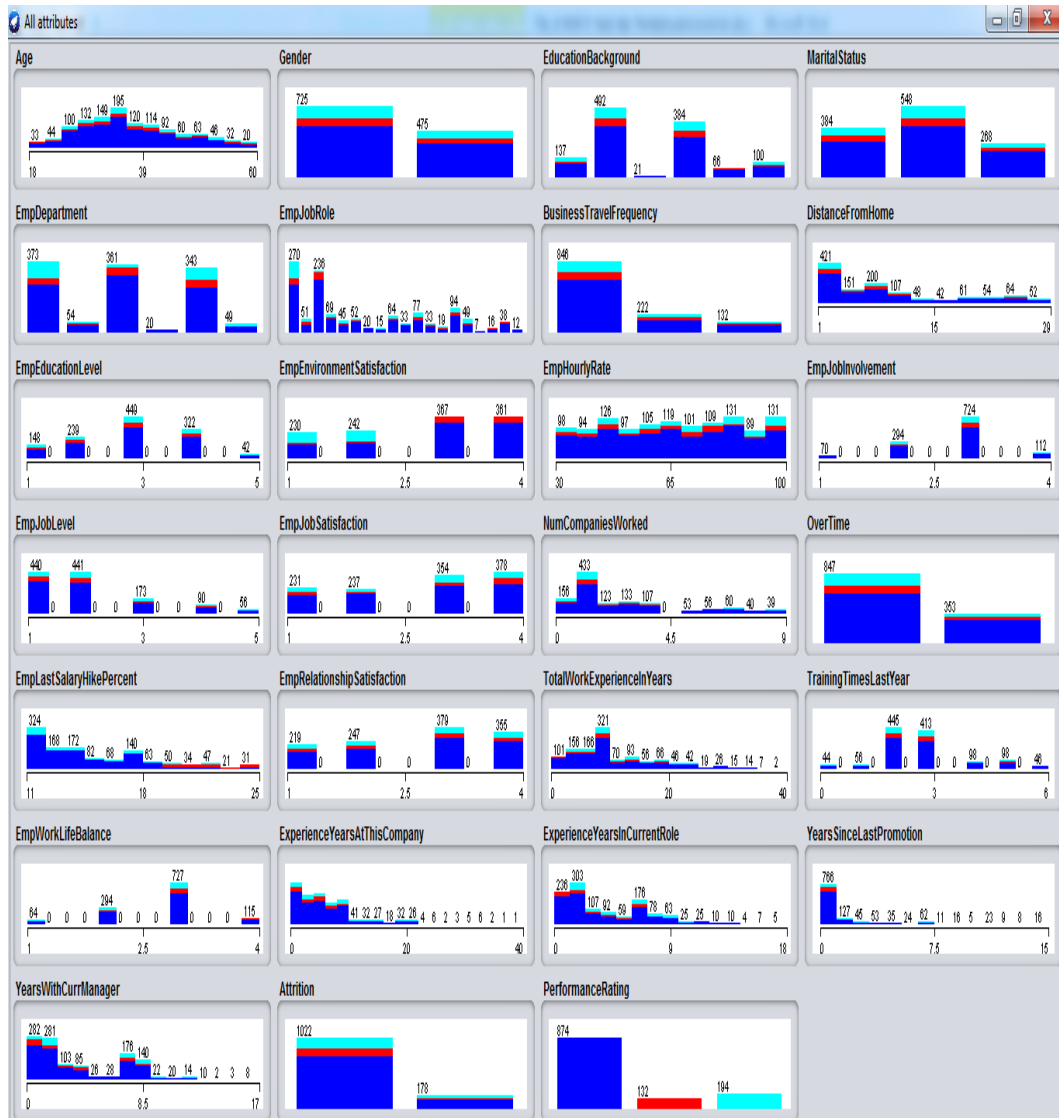
Algoritma	Akurasi	Kappa Statistic	MAE	Precision	Recall	ROC
<b><i>DT J48</i></b>	<b>95,41%</b>	<b>0,8925</b>	<b>0,0432</b>	<b>0,955</b>	<b>0,954</b>	<b>0,964</b>
<i>Random Forest</i>	95%	0,8794	0,0985	0,951	0,950	0,986
<i>Naive Bayes</i>	80,16%	0,5096	0,1639	0,792	0,802	0,866
<i>SVM</i>	82,58%	0,57	0,2644	0,821	0,826	0,791
<i>Logistic</i>	81,75%	0,56	0,1512	0,811	0,818	0,888
<i>KNN</i>	87%	0,6933	0,0874	0,868	0,870	0,853

Hasil kinerja pengujian klasifikasi dengan teknik *Resample* pada gambar 4.1. Menunjukkan nilai dari kinerja teknik *resample* dengan beberapa algoritma yaitu *Desicion Tree J48*, *Random Forest*, *Naive Bayes*, *KNN*, *Logistic* dan *SVM* menunjukkan hasil akurasi terbaik yaitu algoritma ***Decision Tree J48*** dengan nilai akurasi sebesar **95,41%**.



Gambar IV.9.Grafik Hasil Kinerja

Hasil Kinerja Pengujian Klasifikasi *Desicion Tree J48*, *Random Forest*, *Naive Bayes*, *KNN*, *Logistic* dan *SVM* dengan *Resample*



Gambar IV.10. Visualize All Atribut Dataset Asli Tanpa Resample





Gambar IV.11 Visualize All Atribut Dataset Asli + Resample

### 4.3.3 Pengujian klasifikasi dengan Resample dan Feature Selection Algoritm

Tabel IV.17

Rank Attributes

0.32648	17	EmpLastSalaryHikePercent
0.3014	10	EmpEnvironmentSatisfaction
0.149	24	YearsSinceLastPromotion

0.1307	23	ExperienceYearsInCurrentRole
0.12278	25	YearsWithCurrManager
0.09055	5	EmpDepartment
0.07032	22	ExperienceYearsAtThisCompany
0.05964	6	EmpJobRole
0.05723	21	EmpWorkLifeBalance
0.05131	9	EmpEducationLevel
0.04865	16	OverTime
0.04368	13	EmpJobLevel
0.04188	18	EmpRelationshipSatisfaction
0.04009	14	EmpJobSatisfaction
0.03873	26	Attrition
0.03827	20	TrainingTimesLastYear
0.03801	4	MaritalStatus
0.03622	19	TotalWorkExperienceInYears
0.035	3	EducationBackground
0.02553	15	NumCompaniesWorked
0.02353	7	BusinessTravelFrequency
0.01902	8	DistanceFromHome
0.0177	1	Age
0.01611	12	EmpJobInvolvement
0.01476	11	EmpHourlyRate
0.00854	2	Gender
0.01476	11	EmpHourlyRate
0.00854	2	Gender

Berdasarkan hasil dari pengujian menggunakan *correlation attribute eval* didapatkan beberapa atribut teratas yang paling berpengaruh dalam memprediksi peringkat kinerja karyawan yaitu : 1. EmpLastSalaryHikePercent, 2. EmpEnvironmentSatisfaction, 3. YearsSinceLastPromotion, 4. ExperienceYearsInCurrentRole.

## BAB V

### PENUTUP

#### 5.1. Kesimpulan

Penelitian ini dibuat untuk menguji dataset HRM menggunakan teknik resample dengan algoritma C.45 (J48), Support Vector Machine, Naive Bayes, Logistic, K-Nearest Neighbor dan Random Forest.

Kesimpulan dari penelitian ini adalah :

1. Melakukan praprocessing data dengan pengujian dataset dari beberapa algoritma klasifikasi diantaranya J48, Support Vector Machine, Naive Bayes, Logistic, K-Nearest Neighbor dan Random Forest. Kemudian dari beberapa algoritma yang diusulkan akan dievaluasi menggunakan confusion matrix.
2. Dari hasil pengujian dan perbandingan algoritma C.45 (J48), Support Vector Machine, Naive Bayes, Logistic, K-Nearest Neighbor dan Random Forest pada dataset HRM menunjukkan hasil akurasi terbaik yaitu algoritma C.45 (J48) sebesar **92,83%**, Support Vector Machine sebesar **82,33%**, Naive Bayes sebesar **79,16%**, Logistic sebesar **81,25%**, K-Nearest Neighbor sebesar **66,91%** dan Random Forest sebesar **90,41%**.
3. Dari hasil pengujian klasifikasi menggunakan teknik resample pada dataset HRM menggunakan algoritma *Decision Tree J48, Random Forest, Naive Bayes, KNN, Logistic dan SVM* menunjukkan hasil akurasi terbaik yaitu algoritma Decision Tree J48 dengan nilai akurasi sebesar **95,41%**, nilai kappa sebesar **0,8925**, nilai MAE sebesar **0,0432**, Nilai Precision sebesar **0,955**, Nilai Recall sebesar **0,954** dan nilai ROC sebesar **0,964**.
4. Berdasarkan hasil dari pengujian menggunakan correlation attribute eval didapatkanlah beberapa atribut teratas yang paling berpengaruh dalam memprediksi peringkat kinerja karyawan yaitu :

- EmpLastSalaryHikePercent
- EmpEnvironmentSatisfaction
- YearsSinceLastPromotion
- ExperienceYearsInCurrentRole

## **5.2. Saran**

Agar penelitian ini bisa ditingkatkan dan dikembangkan berikut saran-saran yang diusulkan :

1. Membandingkan penelitian dengan dataset yang berbeda
2. Penelitian ini dapat dikembangkan dengan membandingkan algoritma data mining yang lainnya juga tekniknya disesuaikan terlebih dahulu dengan datasetnya agar mendapatkan peningkatan nilai akurasi dan lainnya.
3. Jika mungkin dapat dikembangkan dengan dibuatkan aplikasi dalam memprediksi peringkat kinerja karyawan dengan bobot atribut yang sangat berpengaruh.

## DAFTAR PUSTAKA

- [1] H. Dhika and F. Destiawati, “Penerapan Algoritma C45 Untuk Penilaian Karyawan Pada Restoran Cepat Saji,” no. September, pp. 55–59, 2018, doi: 10.31227/osf.io/zcsfm.
- [2] W. T. Ina *et al.*, “Klasifikasi Tingkat Kelulusan Mahasiswa Prodi Teknik Elektro,” *Semin. Nas. Sains Dan Tek. Fst Undana*, pp. 355–361, 2019.
- [3] I. Handayani, “Algoritma, Enkripsi , Deskripsi , DES dan RSA UNTUK KEAMANAN DATA,” *Jaisek*, vol. 1, no. 2, pp. 89–97, 2019, doi: 10.12928/JASIEK.v13i2.xxxx.
- [4] D. Nofriansyah, K. Erwansyah, and M. Ramadhan, “Penerapan Data Mining dengan Algoritma Naive Bayes Clasifier untuk Mengetahui Minat Beli Pelanggan terhadap Kartu Internet XL ( Studi Kasus di CV. Sumber Utama Telekomunikasi),” *J. Saintikom*, vol. 15, no. 2, pp. 81–92, 2016.
- [5] S. Hilda Kusumahadi, H. Junaedi, and J. Santoso, “Klasifikasi Helpdesk Menggunakan Metode Support Vector Machine,” *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 54–60, 2019, doi: 10.30591/jpit.v4i1.1125.
- [6] C. Java, D. S. Wisdayani, I. M. Nur, R. Wasono, and U. M. Semarang, “Penerapan Algoritma K-Nearest Neighbor dalam Klasifikasi Tingkat Keparahan Korban Kecelakaan Lalu Lintas di Kabupaten Jawa Tengah,” pp. 373–380.
- [7] D. Widiastuti, J. S. Informasi, and U. Gunadarma, “Analisa Perbandingan Algoritma Svm , Naive Bayes , Dan Decision Tree Dalam Mengklasifikasikan Serangan ( Attacks ),” pp. 1–8.
- [8] S. Saifullah, M. Zarlis, Z. Zakaria, and R. W. Sembiring, “Analisa

- Terhadap Perbandingan Algoritma Decision Tree Dengan Algoritma Random Tree Untuk Pre-Processing Data,” *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 1, no. 2, p. 180, 2017, doi: 10.30645/j-sakti.v1i2.41.
- [9] D. I. Baihaqi, A. N. Handayani, and U. Pujiyanto, “Perbandingan Metode Naïve Bayes Dan C4.5 Untuk Memprediksi Mortalitas Pada Peternakan Ayam Broiler,” *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 10, no. 1, pp. 383–390, 2019, doi: 10.24176/simet.v10i1.2846.
- [10] B. Utami and P. Aliandu, “KLASIFIKASI PENENTUAN TIM UTAMA OLAHRAGA HOCKEY MENGGUNAKAN ALGORITMA C4.pdf,” *Proc. Int. Conf. Information, Commun. Technol. Syst.*, vol. 5, no. 4, pp. 1–5, 2013.
- [11] O. Heranova, “Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 443–450, 2019, doi: 10.29207/resti.v3i3.1275.
- [12] A. Rachmat and Y. Lukito, “SENTIPOL: Dataset Sentimen Komentar Pada Kampanye PEMILU Presiden Indonesia 2014 dari Facebook Page,” *Konf. Nas. Teknol. Inf. dan Komun. 2017*, no. December, pp. 218–228, 2016.
- [13] R. R. Pratama, “Analisis Model Machine Learning Terhadap Pengenalan Aktivitas Manusia,” *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 19, no. 2, pp. 302–311, 2020, doi: 10.30812/matrik.v19i2.688.
- [14] E. S. Mona Nasr Ahmed Samir, “A Proposed Model for Predicting Employees’ Performance Using Data Mining Techniques: Egyptian Case Study,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 17, no. 1, pp. 31–40, 2019.

- [15] S. Mulyati, Y. Yulianti, and A. Saifudin, “Penerapan Resampling dan Adaboost untuk Penanganan Masalah Ketidakseimbangan Kelas Berbasis Naïve Bayes pada Prediksi Churn Pelanggan,” *J. Inform. Univ. Pamulang*, vol. 2, no. 4, p. 190, 2017, doi: 10.32493/informatika.v2i4.1440.



## DAFTAR RIWAYAT HIDUP

### A. Biodata Mahasiswa

N.I.M : 14002244  
Nama Lengkap : Rizky Ade Safitri  
Tempat & Tanggal Lahir : Sintang, 24 Januari 1997  
Alamat Lengkap : Jl. Tanjung Raya II Komplek Villa Sejahtera 2  
Jalur Sejahtera 1 Blok J No 9, Kel. Parit Mayor  
Kec. Pontianak Timur, Kalimantan Barat  
No Handphone : 0898-6122-885  
E-mail : [rizkyadesafitri@gmail.com](mailto:rizkyadesafitri@gmail.com)  
Kewarganegaraan : Indonesia  
Status : Belum Menikah

### B. Riwayat Pendidikan Formal

1. SD Negeri 28 Pontianak Utara lulusan tahun 2008
2. SMP Negeri 02 Sui Ambawang tahun 2011
3. SMK Negeri 07 Pontianak lulusan tahun 2014
4. AMIK BSI Pontianak lulusan tahun 2017
5. Universitas BSI Bandung, lulusan tahun 2018

### C. Riwayat Pengalaman Berorganisasi / Pekerjaan

1. Ketua Osis SMKN 7 Pontianak tahun 2011-2012
2. Sekretaris Senat Mahasiswa (SEMA) AMIK BSI Pontianak 2014-2015

### D. Pengalaman Kerja

1. Magang pada bagian perpustakaan AMIK BSI Pontianak
2. Asisten Instruktur AMIK BSI Pontianak
3. Magang pada bagian Administrasi di UBSI Salemba 22 Jakarta Pusat

Jakarta, 06 Agustus 2020



Rizky Ade Safitri