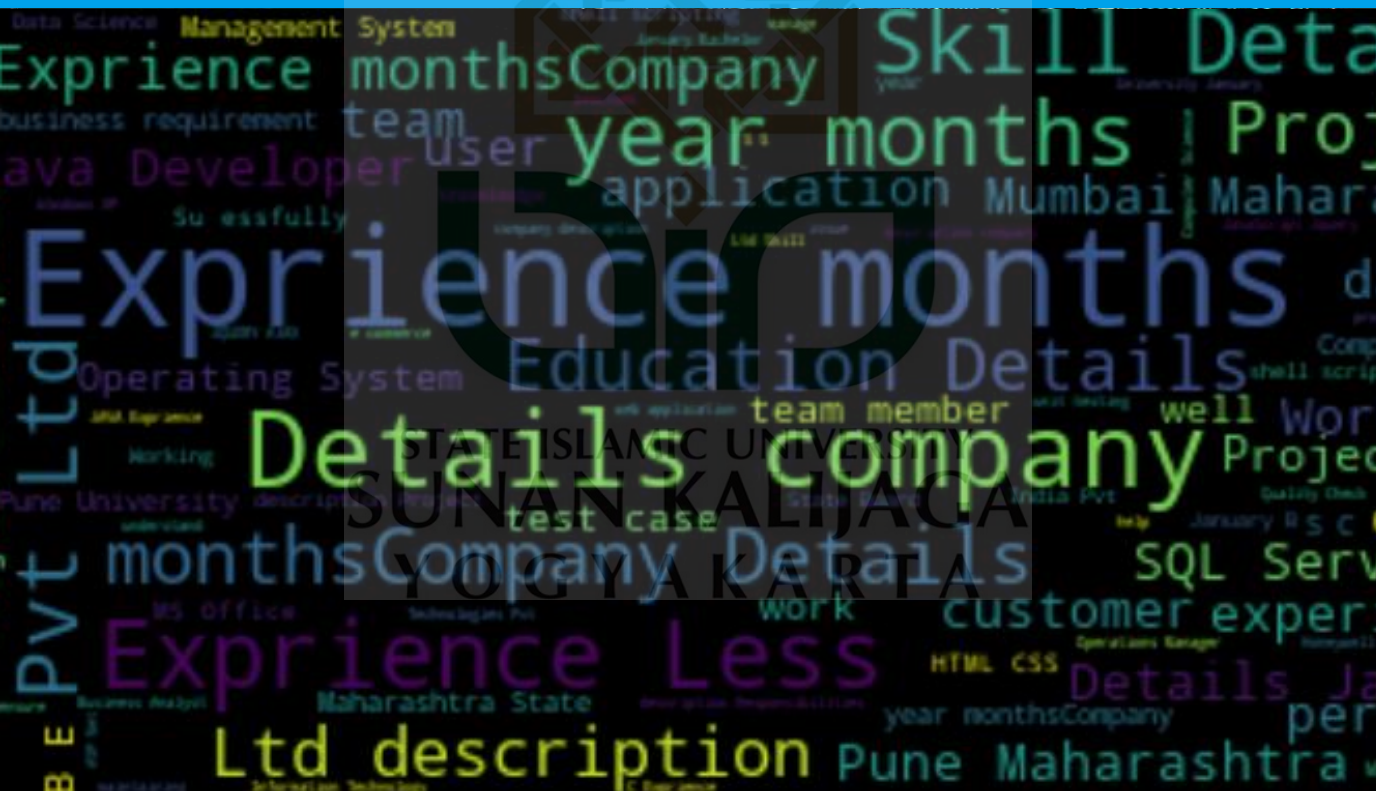


**MODEL OTOMATISASI SELEKSI RESUME LAMARAN KERJA
MENGUNAKAN NATURAL LANGUAGE PROCESSING (NLP)
DAN MACHINE LEARNING (ML)**



**MODEL OTOMATISASI SELEKSI RESUME LAMARAN KERJA
MENGUNAKAN NATURAL LANGUAGE PROCESSING (NLP)
DAN MACHINE LEARNING (ML)**



OLEH:

AFDI FAUZUL BAHAR

22206051011

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA
PROGRAM MAGISTER INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UIN SUNAN KALIJAGA

YOGYAKARTA

2024

PERNYATAAN KEASLIAN

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama : Afdi Fauzul Bahar

NIM : 22206051011

Jenjang : Magister

Program Studi : Informatika

menyatakan bahwa naskah tesis ini secara keseluruhan adalah hasil penelitian/karya saya sendiri, kecuali pada bagian-bagian yang dirujuk sumbernya.

Yogyakarta, 20 Agustus 2024

Saya yang menyatakan,



Afdi Fauzul Bahar

NIM: 22206051011

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

PERNYATAAN BEBAS PLAGIASI

PERNYATAAN BEBAS PLAGIASI

Yang bertanda tangan di bawah ini:

Nama : Afdi Fauzul Bahar

NIM : 22206051011

Jenjang : Magister

Program Studi : Informatika

menyatakan bahwa naskah tesis ini secara keseluruhan benar-benar bebas dari plagiasi. Jika di kemudian hari terbukti melakukan plagiasi, maka saya siap ditindak sesuai ketentuan hukum yang berlaku

Yogyakarta, 20 Agustus 2024

Saya yang menyatakan,

A handwritten signature in black ink, reading "Afdi Fauzul Bahar", is written over a rectangular postage stamp. The stamp is yellow and red, featuring the Garuda Pancasila emblem and the text "5000" and "METAL TEMBEL". The serial number "787ALX338137517" is visible at the bottom of the stamp.

Afdi Fauzul Bahar

NIM: 22206051011

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

HALAMAN PENGESAHAN



KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI

Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

PENGESAHAN TUGAS AKHIR

Nomor : B-1578/Un.02/DST/PP.00.9/08/2024

Tugas Akhir dengan judul : Model Otomatisasi Selesi Resume Lamaran Kerja Menggunakan Natural Language Processing (NLP) dan Machine Learning (ML)

yang dipersiapkan dan disusun oleh:

Nama : AFDI FAUZUL BAHAR, S.Kom.
Nomor Induk Mahasiswa : 22206051011
Telah diujikan pada : Rabu, 21 Agustus 2024
Nilai ujian Tugas Akhir : A-

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

TIM UJIAN TUGAS AKHIR



Ketua Sidang

Prof. Dr. Ir. Shofwatul Uyun, S.T., M.Kom., IPM., ASEAN Eng.
SIGNED

Valid ID: 66ebc3e7e4645



Penguji I

Ir. Maria Ulfah Siregar, S.Kom., MIT., Ph.D.
SIGNED

Valid ID: 66caafc5de612



Penguji II

Dr. Ir. Bambang Sugiantoro, M.T., IPU.,
ASEAN Eng.
SIGNED

Valid ID: 66c7dbcc61b41



Yogyakarta, 21 Agustus 2024
UIN Sunan Kalijaga
Dekan Fakultas Sains dan Teknologi

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

Valid ID: 66cc0376db8c5

PERSETUJUAN TIM PENGUJI UJIAN TESIS



KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI

Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

BERITA ACARA UJIAN TUGAS AKHIR

Penyelenggaraan Ujian Tugas Akhir Mahasiswa

A. Waktu, Tempat dan Status Ujian Tugas Akhir:

1. Hari dan Tanggal : Rabu, 21 Agustus 2024
2. Pukul : 08:00 s/d 09:00 WIB
3. Tempat : FST-1-101
4. Status : Utama

B. Susunan Tim Ujian Tugas Akhir:

No.	Jabatan	Nama	Tanda Tangan
1.	Ketua Sidang	Prof. Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom., IPM., ASEAN Eng.	 Valid ID: 66c6b5c3a9474
2.	Penguji I	Ir. Maria Ulfah Siregar, S.Kom., MIT., Ph.D.	 Valid ID: 66c6b436b286
3.	Penguji II	Dr. Ir. Bambang Sugiantoro, M.T., IPU., ASEAN Eng.	 Valid ID: 66c551781693e

C. Identitas Mahasiswa yang diuji:

1. Nama : AFDI FAUZUL BAHAR, S.Kom.
2. Nomor Induk Mahasiswa : 22206051011
3. Program Studi : Informatika
4. Semester : IV
5. Program : S2
6. Status Kehadiran Mahasiswa : Menghadiri Ujian

D. Judul Tugas Akhir : Model Otomatisasi Selesi Resume Lamaran Kerja Menggunakan Natural Language Processing (NLP) dan Machine Learning (ML)

E. Pembimbing/Promotor:

1. Prof. Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom., IPM., ASEAN Eng.

F. Keputusan Sidang

1. LULUS dengan Perbaikan
2. Predikat Kelulusan : 93.00 (A-)
3. Konsultasi Perbaikan a. _____

b. _____



Valid ID: 66c6b5c3a9474

Yogyakarta, 21 Agustus 2024

Ketua Sidang/Pembimbing/Promotor,

Prof. Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom.,
IPM., ASEAN Eng.

SIGNED

NOTA DINAS PEMBIMBING

NOTA DINAS PEMBIMBING

Yth,
Dekan Fakultas Sains dan Teknologi
UIN Sunan Kalijaga Yogyakarta

Assalamu'alaikum Wr. Wb.

Setelah melakukan bimbingan, arahan, dan koreksi terhadap penulisan tesis yang berjudul:

**MODEL OTOMATISASI SELEKSI RESUME LAMARAN KERJA MENGGUNAKAN
NATURAL LANGUAGE PROCESSING (NLP) DAN MACHINE LEARNING (ML)**

Yang ditulis oleh:

Nama : Afdi Fauzul Bahar
NIM : 22206051011
Jenjang : Magister
Program Studi : Informatika

Saya berpendapat bahwa tesis tersebut sudah dapat diajukan kepada Magister Informatika UIN Sunan Kalijaga untuk diujikan dalam rangka memperoleh gelar Magister Informatika

Wassalamu'alaikum wr wb

Yogyakarta, 20 Agustus 2024

Pembimbing,



Prof. Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom., IPM.,
ASEAN Eng.

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

ABSTRAK

Dalam konteks perekrutan tenaga kerja, proses seleksi resume sering kali memerlukan evaluasi yang cermat dan efisien guna menangani volume lamaran yang besar. Penelitian ini bertujuan untuk mengetahui algoritma terbaik pada proses otomatisasi seleksi resume lamaran kerja menggunakan dataset publik.

Penelitian ini membandingkan kinerja tiga algoritma machine learning: *Random Forest*, *Naive Bayes*, dan *Logistic Regression* dalam otomatisasi seleksi resume menggunakan *Natural Language Processing* (NLP) dan *Machine Learning* (ML). Empat skenario pembagian dataset digunakan untuk mengevaluasi kinerja algoritma. Proses preprocessing mencakup penghapusan tanda baca, tokenisasi, *stemming*, *lemmatization*, dan *TF-IDF vectorization*.

Hasil penelitian menunjukkan bahwa *random forest* memiliki kinerja terbaik dengan akurasi dan *F1-Score* tinggi di semua skenario tanpa indikasi *overfitting*. Model *random forest* terbaik mencapai akurasi 99,48%, *F1-Score* 99,49%, dan *cross-validation score* 99,61%. *Logistic Regression* juga berkinerja sangat baik, dengan akurasi sempurna di beberapa skenario, tetapi dengan risiko *overfitting*. *Naive Bayes*, meskipun cepat dan efisien, menunjukkan kinerja lebih rendah dengan akurasi 92,23%, *F1-Score* 91,35%, dan *cross-validation score* 87,00%. Kesimpulannya, *Random Forest* sebaiknya digunakan karena kinerja konsisten dan kemampuan generalisasi yang kuat. *Logistic Regression* cocok jika interpretabilitas model penting, dengan tambahan regularisasi untuk mengurangi *overfitting*. *Naive Bayes* lebih cocok untuk situasi yang memerlukan kecepatan dan efisiensi.

Kata kunci: NLP, Machine Learning, Random Forest, Naive Bayes, Logistic Regression, Seleksi Resume

ABSTRACT

In the context of workforce recruitment, the resume selection process often requires careful and efficient evaluation to handle a large volume of applications. This study aims to identify the best algorithm for automating the resume selection process using a public dataset.

The research compares the performance of three machine learning algorithms: Random Forest, Naive Bayes, and Logistic Regression in automating resume selection using Natural Language Processing (NLP) and Machine Learning (ML). Four dataset splitting scenarios are used to evaluate the algorithm's performance. The preprocessing steps include punctuation removal, tokenization, stemming, lemmatization, and TF-IDF vectorization.

The results indicate that Random Forest performs the best with high accuracy and F1-Score across all scenarios without signs of overfitting. The best model achieved an accuracy of 99.48%, an F1-Score of 99.49%, and a cross-validation score of 99.61%. Logistic Regression also performed very well, with perfect accuracy in several scenarios but with a risk of overfitting. Naive Bayes, while fast and efficient, showed lower performance with an accuracy of 92.23%, an F1-Score of 91.35%, and a cross-validation score of 87.00%. In conclusion, Random Forest is recommended due to its consistent performance and strong generalization capability. Logistic Regression is suitable when model interpretability is important, with additional regularization to reduce overfitting. Naive Bayes is more appropriate for situations requiring speed and efficiency.

Keywords: *NLP, Machine Learning, Random Forest, Naive Bayes, Logistic Regression, Resume Selection*

MOTTO

يَرْفَعُ اللَّهُ الَّذِينَ ءَامَنُوا مِنْكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ ۗ وَاللَّهُ بِمَا تَعْمَلُونَ خَبِيرٌ

“Allah akan meninggikan orang-orang yang beriman di antaramu dan orang-orang yang diberi ilmu pengetahuan beberapa derajat. Dan Allah Maha Mengetahui apa yang kamu kerjakan. (QS. Al Mujadalah ayat: 11)”

“Jika Kamu tidak sanggup menahan lelahnya belajar maka kamu harus sanggup menahan perihnya kebodohan”



STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

KATA PENGANTAR

Bissmillahirrahmanirrahim,

Assalamu 'alaikum warahmatullahi wabarakatuh

Puji dan syukur kita panjatkan kepada Allah SWT atas nikmat dan hidayah-Nya serta telah memberikan kita kekuatan, kesehatan dan kesabaran. Sehingga penulis dapat menyelesaikan tesis dengan judul “**MODEL OTOMATISASI SELEKSI RESUME LAMARAN KERJA MENGGUNAKAN NATURAL LANGUAGE PROCESSING (NLP) DAN MACHINE LEARNING (ML)**”.

Dalam penyusunan dan penyelesaian karya ini, penulis mengakui banyak mendapat dorongan dan bimbingan dari berbagai pihak. Oleh karena itu, penulis ingin menggunakan kesempatan ini untuk mengucapkan terima kasih kepada:

1. Bapak Pardi dan Ibu Afsoni selaku orang tua.
2. Ibu Dr. Khurul Wardati, M,Si, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sunan Kalijaga.
3. Bapak Dr. Ir. Bambang Sugiantoro, S.Si., M.T., IPM., ASEAN Eng. selaku Kaprodi Magister Informatika sekaligus dosen penguji.
4. Ibu Prof. Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom., IPM., ASEAN Eng., selaku dosen pembimbing tesis
5. Ibu Ir. Maria Ulfah Siregar, S.Kom., MIT., Ph.D., selaku dosen penguji sekaligus dosen pembimbing akademik.
6. Seluruh kawan-kawan Magister Informatika Angkatan 2022.

Penulis juga mengucapkan terima kasih kepada seluruh dosen program Magister Informatika yang telah memberikan banyak ilmu dan wawasan bermanfaat selama pendidikan. Semangat penulis dalam menyelesaikan

naskah ini tidak lepas dari dukungan keluarga dan rekan-rekan. Oleh karena itu, penulis juga ingin memanfaatkan kesempatan ini untuk mengucapkan terima kasih kepada seluruh kerabat, terutama orang tua, yang telah mendoakan dan memotivasi hingga menyelesaikan tesis ini dengan baik.

Akhir kata, naskah ini belum sempurna, oleh karena itu penulis sangat menghargai kritik dan saran yang bersifat membangun. Semoga Allah meridhoi langkah kita, Armin. Oleh karena itu, penulis menyusun tesis ini. Semoga tesis ini dapat dikembangkan dan digunakan sesuai kemampuan masing-masing dan semoga Allah SWT selalu memberikan ilmu yang bermanfaat kepada semua orang dan membimbing mereka di jalan Allah.

Wassalamu'alaikum warahmatullahi wabarakatuh.

Yogyakarta, 20 Agustus 2024



Penulis

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN KEASLIAN.....	ii
PERNYATAAN BEBAS PLAGIASI	iii
HALAMAN PENGESAHAN.....	iv
PERSETUJUAN TIM PENGUJI UJIAN TESIS	v
NOTA DINAS PEMBIMBING	vi
ABSTRAK	vii
<i>ABSTRACT</i>	viii
MOTTO	ix
KATA PENGANTAR.....	x
DAFTAR ISI.....	xii
DAFTAR GAMBAR.....	xv
DAFTAR TABEL	xvi
DAFTAR RUMUS	xvii
BAB I PENDAHULUAN.....	1
1. 1. Latar Belakang.....	1
1. 2. Rumusan Masalah	3
1. 3. Batasan Masalah	3
1. 4. Tujuan Penelitian.....	3
1. 5. Manfaat Penelitian.....	4
1. 6. Keaslian Penelitian.....	4
BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI.....	5
2. 1. Tinjauan Pustaka.....	5
2. 2. Landasan Teori	9
2. 2. 1. Kecerdasan Buatan	9
2. 2. 2. Machine Learning	12
2. 2. 3. Logistic Regression.....	12

2. 2. 4. Naive Bayes	14
2. 2. 5. Random Forest	14
2. 2. 6. Natural Language Processing.....	16
2. 2. 7. Evaluasi Model.....	19
2. 2. 8. Cross-Validation	20
2. 2. 9. Resume Screening.....	21
BAB III METODE PENELITIAN.....	23
3. 1. Dataset.....	23
3. 2. Data Preprocessing	23
3. 2. 1. Remove Punctutation	24
3. 2. 2. Tokenisasi	24
3. 2. 3. Stemming & Lemmatization	25
3. 2. 4. TF-IDF Vectorization.....	25
3. 3. Pembagian Dataset.....	26
3. 4. Pemodelan.....	26
3. 5. Validasi Model.....	26
3. 6. Pengujian Model	27
3. 7. Evaluasi Matrix.....	27
3. 8. Hasil dan Analisis.....	27
BAB IV HASIL DAN PEMBASAN	28
4. 1. Hasil.....	28
4. 1. 1. Deskripsi Dataset	28
4. 1. 2. Preprocessing Data.....	29
4. 1. 3. Pembagian Dataset	38
4. 1. 4. Hasil Pemodelan.....	39
4. 2. Pembahasan.....	64
BAB V KESIMPULAN	68
5. 1. Kesimpulan.....	68

5. 2. Saran 69
DAFTAR PUSTAKA..... lxx



DAFTAR GAMBAR

Gambar 2.1. Cakupan antara AI, ML, dan NLP	10
Gambar 3.1. Diagram alir penelitian.....	24
Gambar 4.1. Distribusi Kategori Pekerjaan	29
Gambar 4.2. Hasil Tokenisasi, Stemming dan Lemmatization	30
Gambar 4.3. Distribusi Data pada Tiap Skenario	38
Gambar 4.4. Heatmap Skenario Pertama dan Kedua pada Random Forest..	40
Gambar 4.5. Heatmap Skenario Pertama untuk Model Naïve Bayes	43
Gambar 4.6. Heatmap Skenario Pertama & Kedua pada Logistic Regression	44
Gambar 4.7. Perbandingan Akurasi, F1-Score dan CV pada Skenario 1	45
Gambar 4.8. Heatmap Skenario Kedua pada Model Naïve Bayes	48
Gambar 4.9. Perbandingan Akurasi, F1-Score dan CV pada Skenario 2	51
Gambar 4.10. Heatmap Skenario Ketiga dan Keempat pada Random Forest	52
Gambar 4.11. Skenario Ketiga Model Naïve Bayes	54
Gambar 4.12. Skenario Ketiga Model Logistic Regression	57
Gambar 4.13. Skenario 3 Perbandingan Akurasi, F1-Score dan CV	59
Gambar 4.14. Heatmap Skenario Kedua pada Model Naïve Bayes	60
Gambar 4.15. Perbandingan Akurasi, F1-Score dan CV pada Skenario 4 ...	63
Gambar 4.16. Perbandingan model pada berbagai skenario	65

DAFTAR TABEL

Table 2.1. Perbandingan usulan dengan penelitian terdahulu.....	8
Table 4.1. Tahapan Preprocessing Data pada Dokumen 1	36
Table 4.2. Hasil Klasifikasi Percobaan 1 dan 2 pada Model Random Forest	42
Table 4.3. Hasil Klasifikasi Percobaan 1 untuk Model Naïve Bayes	46
Table 4.4. Hasil Klasifikasi Percobaan 1 dan 2 pada Logistic Regression ...	47
Table 4.5. Hasil Klasifikasi Percobaan 2 untuk Model Naïve Bayes	50
Table 4.6. Hasil Klasifikasi Percobaan 3 dan 4 untuk model Random Forest	55
Table 4.7. Hasil Klasifikasi Percobaan 3 untuk model Naïve Bayes.....	56
Table 4.8. Hasil Klasifikasi Percobaan 3 dan 4 model Logistic Regression	58
Table 4.9. Hasil Klasifikasi Percobaan 4 pada model Naïve Bayes	62
Table 4.10. Perbandingan akurasi dengan model terdahulu	66



DAFTAR RUMUS

Rumus (2.1).....	13
Rumus (2.2).....	13
Rumus (2.3).....	14
Rumus (2.4).....	14
Rumus (2.5).....	15
Rumus (2.6).....	18
Rumus (2.7).....	18
Rumus (2.8).....	18
Rumus (2.9).....	19
Rumus (2.10).....	19
Rumus (2.11).....	20
Rumus (2.12).....	20



STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

BAB I

PENDAHULUAN

1. 1. Latar Belakang

Karyawan memiliki peran penting dalam kesuksesan organisasi sebagai unsur vital dalam organisasi (Signal Hire, 2021). Kehadiran karyawan yang berkualitas dapat mempengaruhi efisiensi, efektivitas, perencanaan, produksi barang dan jasa, pengawasan kualitas, pemasaran produk, alokasi sumber daya keuangan, serta penentuan tujuan dan strategi organisasi (Linkedin Business, 2023). Mendapatkan karyawan berkualitas tidaklah mudah dan sederhana (Gusdorf, 2008). Karyawan dalam suatu organisasi merupakan sumber daya yang vital dan hanya dapat diperoleh melalui upaya rekrutmen yang efektif (SourceBreaker, 2023). Penelitian menunjukkan bahwa proses rekrutmen yang baik memerlukan informasi yang akurat dan berkelanjutan untuk mendapatkan calon karyawan yang sesuai dengan kualifikasi organisasi (Dima et al., 2024).

Seiring perkembangan teknologi, metode pencarian kerja telah berubah drastis dari metode konvensional ke metode modern menggunakan platform online (Sills, 2014). Platform online memiliki jangkauan yang lebih luas, memungkinkan perusahaan untuk menarik kandidat dari berbagai lokasi, yang sebelumnya tidak mungkin dilakukan dengan metode konvensional (Hire Quotient, 2022). Dengan jangkauan yang lebih luas ini, perusahaan dapat menerima lebih banyak lamaran, meskipun tidak semua pelamar mungkin memenuhi kualifikasi yang dibutuhkan (SourceBreaker, 2023). Hal ini menimbulkan masalah baru yaitu banyaknya lamaran yang masuk tidak sesuai dengan deskripsi pekerjaan. Untuk menangani masalah

ini, perlu dilakukan penyaringan awal pada resume yang diterima (Ri Na, 2024; Signal Hire, 2021). Model seleksi karyawan sangat penting sebagai langkah dalam memilih karyawan yang memenuhi kualifikasi dan standar kualitas agar tujuan organisasi dapat tercapai (Gusdorf, 2008). Organisasi atau perusahaan dapat melakukan seleksi penerimaan karyawan sebagai pola untuk mendapatkan karyawan berkualitas yang sesuai dengan kualifikasi yang dibutuhkan (LinkedIn Business, 2023). Pengurutan resume secara manual menjadi tantangan besar, terutama ketika harus memilih kandidat yang paling sesuai dalam waktu singkat. Oleh karena itu, berbagai metode dan teknologi telah dikembangkan untuk mengatasi masalah ini, termasuk penggunaan teknik Machine Learning (ML) dan Natural Language Processing (NLP) dalam penyaringan resume (Ri Na, 2024).

Proses penyeleksian resume menggunakan machine learning terdiri dari beberapa tahapan yang penting untuk memastikan efisiensi dan akurasi. Tahap pertama adalah pengumpulan data dari berbagai sumber seperti portal pencarian kerja dan database perusahaan (Arvind Kumar et al., 2021). Selanjutnya, data yang diperoleh diolah melalui pra-pemrosesan untuk membersihkan dan menormalisasi teks serta mengekstraksi fitur penting. Setelah itu, algoritma machine learning diterapkan untuk mengidentifikasi pola dan membuat model prediktif. Model ini kemudian dievaluasi untuk memastikan tingkat akurasi yang tinggi (Arvind Kumar et al., 2021).

Di antara berbagai algoritma machine learning, beberapa metode telah terbukti sangat efektif dalam penyaringan resume (Arvind Kumar et al., 2021; Lalitha et al., 2023; Pimpalkar et al., 2023). Random Forest adalah metode ensemble learning yang menggabungkan beberapa pohon keputusan untuk meningkatkan

akurasi klasifikasi. Metode ini dikenal karena kemampuannya dalam mengatasi overfitting dan memberikan hasil yang robust dalam berbagai jenis dataset (Arvind Kumar et al., 2021). Naive Bayes adalah algoritma probabilistik sederhana yang sering digunakan dalam klasifikasi teks. Algoritma ini efektif untuk pengklasifikasian resume karena mampu menangani data teks dengan baik dan memberikan hasil yang cepat (Pimpalkar et al., 2023).

Logistic Regression adalah metode statistik yang digunakan untuk memodelkan hubungan antara satu variabel dependen biner dan satu atau lebih variabel independen. Meskipun sederhana, metode ini sering digunakan dalam klasifikasi karena kemampuannya untuk memberikan interpretasi yang mudah dan hasil yang akurat (Arvind Kumar et al., 2021; Lalitha et al., 2023; Pimpalkar et al., 2023).

1. 2. Rumusan Masalah

Berdasarkan uraian dari latar belakang, rumusan masalah pada penelitian ini adalah bagaimana mendapatkan model terbaik untuk otomatisasi seleksi resume lamaran kerja dengan memanfaatkan *Natural Language Processing* dan *Machine Learning*. Penelitian ini akan membandingkan akurasi, *f1-score* dan *cross-validation* sebagai alat ukur untuk mendapatkan algoritma klasifikasi terbaik.

1. 3. Batasan Masalah

Batasan masalah pada penelitian ini adalah penggunaan algoritma machine learning *naïve bayes*, *logistic regression* dan *random forest* dengan menggunakan dataset publik dari website Kaggle yang memuat 962 data berupa resume dengan 25 kategori.

1. 4. Tujuan Penelitian

Berdasarkan rumusan masalah dan batasan masalah tersebut, tujuan dari penelitian ini adalah untuk mengetahui algoritma terbaik

pada proses otomatisasi seleksi resume lamaran kerja menggunakan dataset publik. Penelitian ini akan membandingkan performa dari algoritma machine learning *naïve bayes*, *logistic regression* dan *random forest* dengan dataset publik yang telah dipre-processing menggunakan *natural language processing*.

1. 5. Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut.

1. Bagi pengembang sistem, dapat dimanfaatkan untuk pemilihan algoritma dalam pengembangan.
2. Bagi pelaku pengguna sistem informasi, dapat termudahkan dengan model yang dikembangkan.
3. Meminimalisir terjadinya *human-error* pada saat seleksi lamaran pekerjaan.

1. 6. Keaslian Penelitian

Penelitian yang berkaitan dengan perbandingan algoritma *random forest*, *naive bayes* dan *logistic regression* pada otomatisasi seleksi resume lamaran kerja menggunakan *natural language processing* (NLP) dan *machine learning* (ML) sejauh pengamatan yang telah dilakukan dengan mempertimbangkan tinjauan pustaka, penelitian tersebut belum pernah dilakukan.

BAB V

KESIMPULAN

5. 1. Kesimpulan

Berdasarkan hasil analisis perbandingan model pada point 4.5, Random Forest adalah model terbaik untuk tugas penyaringan resume, menunjukkan kinerja yang sangat baik dan konsisten di semua skenario tanpa indikasi *overfitting*. Model Random Forest terbaik berdasarkan data yang diberikan adalah yang menggunakan pembagian data 70% latih, 10% validasi, dan 20% uji. Model ini memiliki akurasi 99,48%, F1-Score 99,49%, dan cross-validation 99,61%, yang menunjukkan kemampuan generalisasi yang sangat baik tanpa tanda-tanda *overfitting*.

Logistic Regression juga menunjukkan kinerja yang sangat baik, namun perlu perhatian terhadap risiko *overfitting*. Model Logistic Regression terbaik berdasarkan data yang diberikan adalah yang menggunakan pembagian data 70% latih, 20% validasi, dan 10% uji atau 80% latih, 10% validasi, dan 10% uji. Kedua skenario ini menunjukkan akurasi dan F1-Score sempurna (100,00%) dengan cross-validation yang sangat tinggi (99,31%) namun tidak melebihi dari *random forest*.

Naive Bayes, meskipun efektif dalam situasi tertentu, menunjukkan kinerja yang lebih rendah dibandingkan kedua model lainnya. Model Naive Bayes terbaik berdasarkan data yang diberikan adalah yang menggunakan pembagian data 60% latih, 20% validasi, dan 20% uji. Model ini memiliki akurasi 93,78%, F1-Score 92,96%, dan cross-validation 85,05%.

Adapun kesimpulan dari penelitian ini adalah sebagai berikut.

1. Random Forest

Sebaiknya digunakan karena menunjukkan kinerja yang konsisten dan sangat baik, serta kemampuan generalisasi yang kuat.

2. Penggunaan Logistic Regression

Dapat digunakan jika interpretabilitas model sangat dibutuhkan, dengan tambahan regularisasi untuk mengurangi risiko overfitting.

3. Penggunaan Naive Bayes

Lebih cocok digunakan dalam situasi di mana kecepatan dan efisiensi diutamakan daripada akurasi tinggi, atau sebagai bagian dari ensemble learning.

4. Perbandingan Algoritma

Random forest terbukti lebih baik dibandingkan algoritma lain dikarenakan dapat menangani overfitting pada setiap skenarionya. Hal ini terjadi karena *random forest* menggabungkan beberapa *decision tree* sehingga dapat meningkatkan akurasi dari model yang dibangun.

5. 2. Saran

Adapun saran yang dapat penulis berikan adalah sebagai berikut.

1. Pengembangan aplikasi screening resume yang telah terintegrasi dengan job portal
2. Penanganan overfitting pada algoritma logistic regression
3. Penggunaan dataset dari jobportal yang ada

DAFTAR PUSTAKA

- Alderham, A. H., & Jaha, E. S. (2022). Comparative Semantic Resume Analysis for Improving Candidate-Career Matching. *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, 313–321.
- Alper, S., Tjosvold, D., & Law, K. S. (2000). CONFLICT MANAGEMENT, EFFICACY, AND PERFORMANCE IN ORGANIZATIONAL TEAMS. *Personnel Psychology*, 53(3), 625–642. <https://doi.org/10.1111/J.1744-6570.2000.TB00216.X>
- Arvind Kumar, S., Khusru Akhtar, A., & Ashwani, K. (2021). Resume Screening Using Natural Language Processing and Machine Learning: A Systematic Review. In P. K. and A. T. Swain Debabala and Pattnaik (Ed.), *Machine Learning and Information Processing* (pp. 207–214). Springer Singapore.
- Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, First Edition*. <https://www.nber.org/books-and-chapters/human-capital-theoretical-and-empirical-analysis-special-reference-education-first-edition>
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. In *Information Science and Statistics* (1st ed.). Springer.
- Black, S. E., & Lynch, L. M. (2001). How to Compete: The Impact of Workplace Practices and Information Technology on Productivity. *The Review of Economics and Statistics*, 83(3), 434–445. <https://doi.org/10.1162/00346530152480081>
- Brahim, G. Ben, Tomar, G. S., & Institute of Electrical and Electronics Engineers Saudi Arabia Section. (2022). *Comparative Semantic Resume Analysis for Improving Candidate-Career Matching*.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324/METRICS>
- Chatman, J. A. (1989). Improving Interactional Organizational Research: A Model of Person-Organization Fit. *The Academy of Management Review*, 14(3), 333. <https://doi.org/10.2307/258171>
- Dima, J., Gilbert, M. H., Dextras-Gauthier, J., & Giraud, L. (2024). The effects of artificial intelligence on human resource activities and the roles of the human resource triad: opportunities and challenges. *Frontiers in Psychology*, 15, 1360401. <https://doi.org/10.3389/FPSYG.2024.1360401/BIBTEX>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>

- Gueutal, H. G., Strohmeier, S., & Kabst, R. (2009). Organizational adoption of e/HRM in Europe: An empirical exploration of major adoption factors. *Journal of Managerial Psychology*, 24(6), 482–501. <https://doi.org/10.1108/02683940910974099/FULL/XML>
- Gusdorf, M. L. (2008). *Recruitment and Selection: Hiring the Right Person A two-part learning module for undergraduate students Staffing MAnAgeMent instructor's Manual*. www.shrm.org/hrededucation
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer. <https://doi.org/https://doi.org/10.1007/978-0-387-84858-7>
- Hire Quotient. (2022). *Traditional vs Modern Hiring | HireQuotient*. <https://www.hirequotient.com/blog/traditional-vs-modern-hiring>
- Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (Vol. 3). <https://web.stanford.edu/~jurafsky/slp3/>
- Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Third Edition draft: Vol. Third Edition*.
- Lalitha, B., Kadiyam, S., Kalidindi, R. V., Vemparala, S. M., Yarlagadda, K., & Chekuri, S. V. (2023). Applicant Screening System Using NLP. *International Conference on Innovative Data Communication Technologies and Application, ICIDCA 2023 - Proceedings*, 379–383. <https://doi.org/10.1109/ICIDCA56705.2023.10099953>
- LinkedIn Business. (2023). *Global Talent Trends | LinkedIn Talent Solutions*. <https://business.linkedin.com/talent-solutions/global-talent-trends#hiring>
- Lokesh. S, Balaje. S, M., Prathish. E, & B. Bharathi. (2022a). Resume Screening and Recommendation System using Machine Learning Approaches. *Computer Science & Engineering: An International Journal*, 12(1), 1–7. <https://doi.org/10.5121/cseij.2022.12101>
- Lokesh. S, Balaje. S, M., Prathish. E, & B. Bharathi. (2022b). Resume Screening and Recommendation System using Machine Learning Approaches. *Computer Science & Engineering: An International Journal*, 12(1), 1–7. <https://doi.org/10.5121/cseij.2022.12101>
- Mankawade, A., Pungliya, V., Bhonsle, R., Pate, S., Purohit, A., & Raut, A. (2023a). Resume Analysis and Job Recommendation. *2023 IEEE 8th International Conference for Convergence in Technology, I2CT 2023*. <https://doi.org/10.1109/I2CT57861.2023.10126171>
- Mankawade, A., Pungliya, V., Bhonsle, R., Pate, S., Purohit, A., & Raut, A. (2023b). Resume Analysis and Job Recommendation. *2023 IEEE 8th International Conference for Convergence in Technology, I2CT 2023*. <https://doi.org/10.1109/I2CT57861.2023.10126171>

- Manning, C. D., Raghavan, P., & Schütze, H. (2009a). *An Introduction to Information Retrieval*. Cambridge University Press.
- Manning, C. D., Raghavan, P., & Schütze, H. (2009b). *An Introduction to Information Retrieval*.
https://web.stanford.edu/~jurafsky/slp3/ed3bookfeb3_2024.pdf
- Mccallum, A., & Nigam, K. (1998). *A Comparison of Event Models for Naive Bayes Text Classification*. www.aaai.org
- Mitchell, T. M. (1997). *Machine Learning*.
- Naga Venkata Sai Raghavendra, M. (n.d.). *RESUME SCREENING USING MACHINE LEARNING*. www.jespublication.com
- Pant, D., Pokhrel, D., & Poudyal, P. (2022). Automatic Software Engineering Position Resume Screening using Natural Language Processing, Word Matching, Character Positioning, and Regex. *Proceedings of the 2022 5th International Conference on Advanced Systems and Emergent Technologies, IC_ASET 2022*, 44–48.
https://doi.org/10.1109/IC_ASET53395.2022.9765916
- Pimpalkar, A., Lalwani, A., Chaudhari, R., Inshall, M., Dalwani, M., & Saluja, T. (2023). Job Applications Selection and Identification: Study of Resumes with Natural Language Processing and Machine Learning. *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2023*.
<https://doi.org/10.1109/SCEECS57921.2023.10063010>
- Raghavendra, M. N. V. S. (2022). RESUME SCREENING USING MACHINE LEARNING. *Journal of Engineering Sciences*, 13(09,2022).
 www.jespublication.com
- Ri Na, S. (2024). Application of Artificial Intelligence in Recruitment and Selection. *Academic Journal of Science and Technology*, 9(2), 56.
- Russell, S. J., & Norvig, P. (2021a). *Artificial Intelligence: A Modern Approach* (Vol. 4). Pearson Education, Inc.
- Russell, S. J., & Norvig, P. (2021b). *Artificial Intelligence: A Modern Approach* (Vol. 4). Pearson Education, Inc.
- Signal Hire. (2021). *Online vs Offline Recruitment Practices: Which is better?* / *Daily Host News*. <https://www.dailyhostnews.com/online-vs-offline-recruitment-practices>
- Sills, M. (2014). *E-recruitment: A comparison with traditional recruitment and the influences of social media*.
- SourceBreaker. (2023). *Traditional Recruitment vs. Modern Recruitment* / *SourceBreaker*. <https://www.sourcebreaker.com/traditional-vs-modern-recruitment/>
- Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, 87(3), 355–374. <https://doi.org/10.2307/1882010>

Upadhyay, A. K., & Khandelwal, K. (2018). Applying artificial intelligence: implications for recruitment. *Strategic HR Review*, 17(5), 255–258. <https://doi.org/10.1108/SHR-07-2018-0051>





PROGRAM STUDI INFORMATIKA
PROGRAM MAGISTER FAKULTAS SAINS DAN TEKNOLOGI
UIN SUNAN KALIJAGA