ANALISIS KEBUTUHAN *SKILL* PASAR KERJA BIDANG TEKNOLOGI INFORMASI JOBSTREET INDONESIA MENGGUNAKAN ALGORITMA *MACHINE LEARNING*

Skripsi

Oleh
MUHAMMAD RIFQI MAJID
NPM 2015061036

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024

ANALISIS KEBUTUHAN *SKILL* PASAR KERJA BIDANG TEKNOLOGI INFORMASI JOBSTREET INDONESIA MENGGUNAKAN ALGORITMA *MACHINE LEARNING*

Oleh

MUHAMMAD RIFQI MAJID

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA TEKNIK

Pada

Program Studi Teknik Informatika
Jurusan Teknik Elektro
Fakultas Teknik

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024

ABSTRAK

ANALISIS KEBUTUHAN *SKILL* PASAR KERJA BIDANG TEKNOLOGI INFORMASI JOBSTREET INDONESIA MENGGUNAKAN ALGORITMA *MACHINE LEARNING*

Oleh

MUHAMMAD RIFQI MAJID

Seiring dengan pesatnya perkembangan teknologi informasi, kebutuhan akan keterampilan di bidang ini semakin meningkat. Situs Jobstreet menyediakan berbagai kualifikasi, termasuk pekerjaan di bidang informasi teknologi. Oleh karena itu, perlu dilakukan klasifikasi untuk mengidentifikasi tren keterampilan. Data lowongan dari Jobstreet dapat dimanfaatkan sebagai data mentah untuk menghasilkan klasifikasi keterampilan IT yang komprehensif. Penelitian ini akan berfokus pada eksplorasi algoritma *machine learning*, dalam konteks klasifikasi untuk menganalisis tren keterampilan. Penelitian ini juga membandingkan akurasi model dalam klasifikasi data, memberi visualisasi hasil model *data mining*, serta mengidentifikasi sub-kategori dan tren keterampilan kerja yang dibutuhkan oleh industri. Dengan menggunakan *framework* pengembangan CRISP-DM, serta model algoritma KNN, NBC, dan SVM. Metode penelitian mencakup pengumpulan data melalui teknik *scraping*, pengolahan data dengan algoritma *machine learning* (tokenisasi, penghapusan stopword, *stemming*, visualisasi n-gram dan *word embeddings*), dan visualisasi data melalui Looker Studio. Hasil penelitian menunjukkan bahwa model SVM unggul dalam akurasi sebesar 86,75%, diikuti KNN dengan akurasi 83,33%, dan NBC dengan akurasi 79,49%. Tren sub-kategori pekerjaan dengan kebutuhan paling banyak, seperti *Business/System Analyst* (34,1%), diikuti oleh *Network & System Administration* (22,6%), dan *Developer/Programmer* (8%). Penelitian ini menunjukkan keunggulan algoritma SVM dibandingkan dengan algoritma lain, menunjukkan bahwa model tersebut memiliki kinerja baik dalam hal klasifikasi teks.

Kata Kunci: Jobstreet, *data mining*, CRISP-DM, keterampilan kerja, klasifikasi

ABSTRACT

ANALYSIS OF SKILL REQUIREMENTS IN THE INFORMATION TECHNOLOGY JOB MARKET ON JOBSTREET INDONESIA USING MACHINE LEARNING ALGORITHMS

By

MUHAMMAD RIFQI MAJID

With the rapid advancement of information technology, the demand for skills in this field is growing significantly. Jobstreet provides various qualifications, including jobs in information technology. Therefore, classification is necessary to identify skill trends. Job vacancy data from Jobstreet can be utilized as raw data to generate a comprehensive classification of IT skills. This research focuses on exploring machine learning algorithms in the context of classification to analyze skill trends. It also compares model accuracy in data classification, provides visualizations of data mining results, and identifies sub-categories and skill trends required by the industry. The study adopts the CRISP-DM framework and employs KNN, NBC, and SVM algorithms. The research methodology includes data collection through scraping techniques, data processing using machine learning algorithms (tokenization, stopword removal, stemming, n-gram visualization, and word embeddings), and data visualization through Looker Studio. The results show that the SVM model excels with an accuracy of 86.75%, followed by KNN at 83.33%, and NBC at 79.49%. The most in-demand job sub-categories include Business/System Analyst (34.1%), Network & System Administration (22.6%), and Developer/Programmer (8%). This study demonstrates the superiority of the SVM algorithm over other algorithms, highlighting its strong performance in text classification tasks.

Keywords: *JobStreet, data mining, CRISP-DM, skills, classification*

Judul Skripsi	:	**ANALISIS KEBUTUHAN *SKILL* PASAR KERJA BIDANG TEKNOLOGI INFORMASI JOBSTREET INDONESIA MENGGUNAKAN ALGORITMA *MACHINE LEARNING***
Nama Mahasiswa	:	**Muhammad Rifqi Majid**
Nomor Pokok Mahasiswa	:	2015061036
Program Studi	:	Teknik Informatika
Fakultas	:	Teknik

MENYETUJUI

1. Komisi Pembimbing

Pembimbing Utama	Pembimbing Pendamping
Ir. Ing. Hery Dian Septama, S.T. IPM.	**Mahendra Pratama, S.T., M.Eng.**
NIP. 19850915200812100	NIP. 199112152019031013

2. Mengetahui

Ketua Jurusan Teknik Elektro	Ketua Program Studi Teknik Informatika
Herlinawati, S.T., M.T.	**Yessi Mulyani, S.T., M.T.**
NIP. 197103141999032001	NIP.197312262000122001

3.

MENGESAHKAN

1. Tim Penguji

Ketua : **Ir. Ing. Hery Dian Septama, S.T. IPM.** .....

Sekretaris : **Mahendra Pratama, S.T., M.Eng.** .....

Penguji : **Mona Arif Muda, S.T., M.T.** .....

2. Dekan Fakultas Teknik

Dr. Eng. Ir. Helmy Fitriawan, S.T., M.Sc.
NIP. 19750928 200112 1 002

Tanggal Lulus Ujian Skripsi : **22 November 2024**

SURA T PERNYATAAN

Saya yang bertandatangan di bawah ini, menyatakan bahwa skripsi saya dengan judul "Analisis Kebutuhan Skill Pasar Kerja Bidang Teknologi Informasi Jobstreet Indonesia Menggunakan Algoritma Machine Learning" dibuat oleh saya sendiri. Semua basil yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila di kemudian hari terbukti bahwa skripsi ini merupakan salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan hukum atau akademik yang berlaku.

Bandar Lampung, 21 November 2024
Pembuat pernyataan

Muhammad Rifqi Majid
NPM. 2015061036

RIWAYAT HIDUP

Penulis bernama Muhammad Rifqi Majid merupakan anak tunggal dari pasangan Bapak Edy Syofian dan Ibu Feri Indrawati. Penulis lahir di Metro, pada tanggal 28 Desember 2001. Penulis menyelesaikan pendidikannya di Taman Kanak-Kanak (TK) Aisyiyah Busthanul Athfal 1 Pringsewu 2008, Pendidikan Sekolah Dasar (SD) Muhammadiyah Pringsewu pada tahun 2014, Sekolah Menengah Pertama (SMP) Negeri 3 Pringsewu pada tahun 2017, dan Sekolah Menengah Atas (SMA) Negeri 2 Pringsewu pada tahun 2020.

Pada tahun 2020 penulis terdaftar sebagai mahasiswa Program Studi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik Universitas Lampung melalui jalur SBMPTN (Seleksi Bersama Masuk Perguruan Tinggi Negeri). Selama menjadi mahasiswa, penulis melakukan beberapa kegiatan, antara lain:

1. Mendapatkan beasiswa pendidikan Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, Kartu Indonesia Pintar Kuliah (KIP Kuliah) selama 4 tahun sejak 2020.
2. Menjadi anggota biasa Himpunan Mahasiswa Teknik Elektro Universitas Lampung, Departemen Pendidikan, dan Pengembangan Diri, Divisi Pendidikan pada tahun 2020.
3. Menjadi anggota biasa Himpunan Mahasiswa Teknik Elektro Universitas Lampung, Departemen Pendidikan, dan Pengembangan, Divisi Kerohanian pada tahun 2022.
4. Menjadi anggota Forum Silaturahmi dan Studi Islam Universitas Lampung (Fossi-FT Unila), Departemen Media Informasi pada tahun 2021.
5. Berhasil menyelesaikan program Kredensial Mikro Mahasiswa Indonesia (KMMI) bidang Teknologi Multimedia yang diselenggarakan oleh LP3M Universitas Lampung pada tahun 2021.

6. Berhasil menyelesaikan program Merdeka Belajar Kampus Merdeka (MBKM) kampus dengan proyek penelitian dosen dengan topik “Pengembangan HUD Mobil Listrik Unila 1 (Electrical Vehicle Unit 1/EVU-1) yang diselenggarkaan oleh Universitas Lampung pada tahun 2022.
7. Berhasil menyelesaikan program Studi Independen Bersertifikat yang diselenggarakan oleh Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi dengan pembelajaran *Java for Android Developer* di PT. Hacktivate Teknologi Indonesia pada tahun 2022.
8. Berhasil menyelesaikan program Baparekraf Developer Day yang diselenggarakan oleh Kementerian Pariwisata dan Ekonomi Kreatif dengan pemebelajaran Profesional Google IT Support pada tahun 2022 tahap 2.
9. Berhasil menyelesaikan program Magang/Kerja Praktik (KP) di Badan Pusat Statistik (BPS) Kabupaten Pringsewu pada tahun 2023.
10. Melaksanakan Kuliah Kerja Nyata di Desa Sri Kuncoro, Kecamatan Semaka, Kabupaten Tanggamus, Provinsi Lampung selama 30 hari terhitung dari bulan Januari sampai dengan Februari 2023.

MOTTO

"Seluruh kejadian yang terjadi adalah skenario Allah, jalani, nikmati, dan syukuri"

(Penulis)

"Kapanpun, dimanapun, dalam kondisi apapun jangan pernah tinggalkan sholat"

(Bapak & Ibu)

"Allah akan meninggikan orang-orang yang beriman diantaramu dan orang-orang yang diberi ilmu pengetahuan beberapa derajat."

(Q.S. Al-Mujadalah : 11)

"Kamu tidak harus menjadi hebat untuk memulai, tetapi kamu harus mulai untuk menjadi hebat."

(Zig Ziglar)

"Jangan pernah merasa bersalah untuk memulai awal baru lagi."

(Rupi Kaur)

PERSEMBAHAN

Bismillaahirrohmaanirrahiim,
Dengan mengharapkan ridho dari Allah SWT,
Kupersembahkan karya skripsiku ini untuk orang-orang yang kusayangi dengan setulus hati:
Orangtua tercinta,
Keluargaku,
Teman-Temanku,
Dan
Orang-orang yang telah membantu hidupku

Terima kasih untuk segalanya,
Kalian adalah hartaku yang paling berharga.

SANWACANA

Puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat dan hidayat-Nya sehingga penulis dapat menyelesaikan penyusunan skripsi/tugas akhir ini dengan judul "Analisis Kebutuhan Skill Pasar Kerja Bidang Teknologi Informasi Jobstreet Indonesia Menggunakan Algoritma *Machine Learning*".

Dalam pelaksanaan dan pembuatan skripsi/tugas Akhir ini penulis menerima dukungan baik secara moril maupun materil yang sangat berharga dari berbagai pihak. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada semua pihak yang telah membantu, khususnya kepada:

1. Kedua orangtua tercinta dan seluruh keluarga penulis yang tidak hentinya mendo'akan serta memberikan dorongan semangat dan materi;
2. Bapak Dr. Eng. Helmy Fitriawan, S.T., M.Sc., selaku Dekan Fakultas Teknik Universitas Lampung;
3. Ibu Herlinawati, S.T., M.T. selaku Ketua Jurusan Teknik Elektro Universitas Lampung;
4. Ibu Yessi Mulyani, S.T., M.T. selaku Ketua Program Studi Teknik Informatika Universitas Lampung dan telah membantu proses kelancaran pengerjaan penelitian;
5. Bapak Ir. Ing. Hery Dian Septama, S.T. IPM. selaku Pembimbing Utama dan Pembimbing Akademik yang selalu meluangkan waktunya untuk memberikan bimbingan dan dukungan serta memudahkan penulis dalam menyelesaikan penelitian ini;
6. Bapak Mahendra Pratama, S.T., M.Eng., selaku Pembimbing Pendamping yang selalu memberikan dukungan serta bimbingan agar menjadi lebih baik;
7. Bapak Mona Arif Muda, S.T., M.T. selaku Penguji yang telah memberikan banyak saran dan masukan terhadap penelitian ini;Mba Asliana Rika selaku

Admin Program Studi Teknik Informatika yang telah banyak membantu penulis dalam segala urusan administrasi selama perkuliahan;

8. Seluruh dosen dan staf Jurusan Teknik Informatika Unila yang memberi masukan dan mempermudah proses pembuatan skripsi/tugas akhir ini;
9. Teman-teman KITA BISA dengan anggota Veni, Era, Kinanti, Habibi, dan Haris menemani dari masa SMA yang telah banyak memberikan hiburan, dukungan, dan membawa ceria di kala gelisah;
10. Keluarga besar Teknik Elektro Angkatan 2020 yang telah menjadi teman seperjuangan sejak mahasiswa baru. Terimakasih telah mewarnai masa perkuliahan penulis dan menulis banyak cerita bersama.

Penulis berharap agar laporan ini dapat menjadi referensi bagi pengembangan keilmuan di bidang teknik informatika. Oleh karena itu, semoga penelitian ini bermanfaat bagi yang membacanya.

Bandar Lampung, 21 November 2023
Penulis,

Muhammad Rifqi Majid

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Gambar **Halaman**

DAFTAR TABEL

Tabel **Halaman**

DAFTAR ISTILAH

Istilah	Deskripsi
API	*Application Programming Interface*
Business Understanding	Tahap pemahaman bisnis
CRISP-DM	*Cross Industry Standart Process for Data Mining*
Database	Basis data
Data Mining	Penambangan data
Data Preparation	Tahap persiapan data
Data Scrapping	Pengambilan data
Data training	Set data pelatihan
Data testing	Set data pengujian
Data Understanding	Tahap pemahaman data
Deployment	Tahap penyebaran
EDA	*Exploratory Data Analysis*
Evaluation	Tahap evaluasi
FP	*False Positive*
Framework	Kerangka kerja
Hyperplane	Fungsi yang dapat digunakan untuk pemisah antar kelas dalam klasifikasi
KNN	K-Nearest Neighbors
Modeling	Tahap pemodelan
N	*Negative*
NBC	*Naïve Bayes Classifier*
P	*Positive*

Skill	Kualifikasi yang dibutuhkan pekerjaan. Berupa hardskill.
Supervised Learning	Teknik dalam pembelajaran mesin yang menggunakan set data yang telah diberi label
SVC	*Support Vector Classifier*
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency-Inverse Document Frequency*
TP	*True Prositive*
TN	*True Negative*
Unsupervised Learning	Teknik dalam pembelajaran mesin yang menggunakan set data tanpa label atau petunjuk manusia
VS Code	*Visual Studio Code*

I. PENDAHULUAN

1.1 Latar Belakang

Dalam era digital yang terus berkembang, teknologi informasi menjadi pilar penting bagi berbagai sektor industri. Solusi teknologi terus berkembang dengan rata-rata 7.2% selama dua belas tahun terakhir, data tersebut sejalan dengan pertumbuhan kebutuhan tenaga ahli bidang informasi teknologi. Perusahaan berlomba-lomba mencari kandidat dengan keterampilan yang relevan untuk menghadapi tantangan pasar yang dinamis. Situs penyedia lowongan kerja seperti Jobstreet memfasilitasi kebutuhan ini dengan menyediakan *platform* pencari kerja [1][2]. Jobstreet sebagai situs terbesar di Asia Tenggara, Jobstreet tidak hanya menawarkan berbagai kategori pekerjaan tetapi juga memberikan kemudahan dalam proses pencocokan kualifikasi antara pelamar dan pemberi kerja, khususnya dalam konteks bidang teknologi informasi (TI) [2]. Seiring meningkatnya kebutuhan akan lowongan tersebut, data lowongan pekerjaan yang tersedia di Jobstreet menjadi sumber informasi yang kaya dan strategis. Dengan data yang tersedia, berbagai tren keterampilan yang diminati perusahaan dapat diidentifikasi secara komprehensif. Model *data mining* diadopsi karena dikenal mampu menggali pola dan pengetahuan tersembunyi dalam data [3][4].

Data mining adalah serangkaian proses penemuan pola dan informasi bermakna dari kumpulan data yang besar [4][5]. Dalam konteks data lowongan pekerjaan, *data mining* memungkinkan identifikasi keterampilan yang sedang tren di pasar kerja teknologi informasi. Teknik ini memiliki berbagai keunggulan, misalkan efisiensi dalam pencarian data, peningkatan perencanaan strategis, perbaikan pengambilan keputusan, dan pengungkapan wawasan baru yang berguna bagi perusahaan maupun pencari kerja. Penelitian ini menyoroti pentingnya teknologi *data mining*

dalam menggali informasi strategis dari data lowongan pekerjaan.

Dalam penelitian ini, teknik klasifikasi dipilih untuk mengidentifikasi tren keterampilan TI berdasarkan data lowongan pekerjaan dari Jobstreet. Algoritma yang digunakan mencakup *k-Nearest Neighbor* (KNN), *Naïve Bayes Classifier* (NBC), dan *Support Vector Machine* (SVM) [6][7][8]. Pemilihan algoritma ini didasarkan pada kemampuannya untuk menangani berbagai jenis data, termasuk teks, serta akurasi yang dapat dibandingkan secara empiris. Framework *Cross-Industry Standard Process for Data Mining* (CRISP-DM) digunakan dalam penelitian ini untuk memberikan struktur yang sistematis. Metode ini memberikan tahapan-tahapan yang terestruktur dalam pengembangan data mining, dalam konteks klasifikasi teks. CRISP-DM terdiri dari enam tahap utama, *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment* [9][10]. Proses pengembangan *data mining* dengan manambahkan teknik *exploratory data analysis* (EDA) pada *data understanding*, guna memberikan hasil analis data yang komprehensif dan terukur [8]. Proses pengumpulan data menggunakan teknik *data scrapping* dengan bahasa pemrograman Python, kemudian proses klasifikasi data dilakukan dengan algoritma klasifikasi KNN, NBC, dan SVM.

1.2 Rumusan Masalah

Berdasarkan permasalahan pada latar belakang maka masalah yang akan dipecahkan melalui penelitian ini yaitu:

1. Seberapa akurat penerapan model algoritma KNN, NBC, dan SVM untuk klasifikasi data?
2. Bagaimana pengaruh penggunaan teknik *machine learning* bagi kualitas akurasi model?
3. Mengetahui sub-kategori dan tren *skill* informasi dan teknologi yang dihasilkan oleh pemodelan *data mining*?

1.3 Tujuan Penelitian

Penelitian ini memiliki tujuan yang tercantum dalam tujuan penelitian, tujuan penelitian ini sebagai berikut:

1. Membuktikan tingkat keakuratan model algoritma KNN, NBC, dan SVM untuk klasifikasi data,
2. Mengetahui pengaruh penggunaan teknik *machine learning* bagi kualitas akurasi model,
3. Mampu menampilkan sub-kategori dan macam-macam *skill* informasi dan teknologi yang dihasilkan oleh pemodelan *data mining*.

1.4 Manfaat Penelitian

Penelitian ini diharapkan mampu memberi manfaat yang tercantum pada manfaat penelitian, manfaat penelitian yang akan diberikan sebagai berikut:

1. Menampilkan seberapa akurat penerapan model algoritma KNN, NBC, dan SVM untuk klasifikasi data,
2. Mengatahui pengaruh penggunaan teknik *machine learning* bagi kualitas akurasi model,
3. Menampilkan sub-kategori dan macam-macam *skill* informasi dan teknologi yang dihasilkan oleh pemodelan *data mining*.

1.5 Batasan Masalah

Penelitian ini dibatasi hanya beberapa poin-poin dasar yang tercantum dalam batasan masalah, batasan masalah dalam penelitian ini sebagai berikut:

1. Proses penelitian *data mining* memanfaatkan *framework* CRISP-DM, dan membandingkan tiga algoritma klasifikasi KNN, NBC, dan SVM,
2. Visualisasi data tahap *deployment* menggunakan *tools* Google Looker Studio,
3. Data yang digunakan dalam penelitian ini diambil dari API situs Jobstreet Indonesia pada Januari, April, dan Juni 2024.

1.6 Sistematika Penulisan Skripsi

Sistematika penulisan skripsi terbagi menjadi 5 (lima) bab untuk memudahkan dalam penulisan, antara lain:

BAB I : PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, tujuan, batasan masalah, manfaat penelitian, dan sistematika penulisan skripsi.

BAB II : TINJAUAN PUSTAKA

Bab ini berisi tentang teori-teori dasar terkait dengan penyusunan laporan penelitian yaitu, jobstreet, *data mining*, klasifikasi, *machine learning*, VS Code, python, looker studio, CRISP-DM, dan penelitian terkait.

BAB III : METODOLOGI PENELITIAN

Bab ini berisi tentang waktu dan tempat penelitian, alat dan bahan penelitian, dan metode penelitian.

BAB IV : HASIL DAN PEMBAHASAN

Bab ini berisi hasil dan pembahasan terkait dengan penyusunan laporan penelitian yaitu, *business understanding*, *data understanding*, *data preparation*, *modelling*, *evalution*, dan *deployment*.

BAB V : KESIMPULAN DAN SARAN

Bab ini memuat kesimpulan berdasarkan hasil pembahasan laporan penelitian, serta berisi saran perbaikan dan pengembangan lebih lanjut.

DAFTAR PUSTAKA

Bab ini memuat daftar sumber kutipan teori - teori yang dijadikan acuan dalam menulis laporan.

LAMPIRAN

Lampiran memuat dokumentasi berkas-berkas penunjang penulisan laporan, berupa foto, dan dokumen lain.

II. TINJAUAN PUSTAKA

2.1 Jobstreet

Jobstreet merupakan situs penyedia layanan informasi lowongan kerja berbasis *online*. Jobstreet didirikan oleh Mark Chang Mun Kee di Malaysia pada tahun 1997. Menurut Forbes 2008, saat ini JobStreet merupakan perusahaan penyedia layanan informasi lowongan kerja daring terbesar di Asia Tenggara [11]. Saat ini Jobstreet telah beroperasi di negara Malaysia, Indonesia, Philippines, dan Singapura [1].

Dalam persaingan bisnis antar perusahaan penyedia layanan lowongan kerja yang sangat beragam JobStreet mampu bersaing dengan beberapa keunggulan dibandingkan kompetitor lain. Keunggulan yang ditawarkan oleh JobStreet antara lain [2]:

1. Memiliki basis data kandidat kerja terbesar di Asia,
2. Wawasan yang relevan, tepat, dan realistis,
3. Tingkat kecocokan kandidat yang lebih tinggi,
4. Solusi lengkap untuk perekrutan,
5. Dukungan berkualitas dari mitra JobStreet yang berpengalaman.

2.2 Data Mining

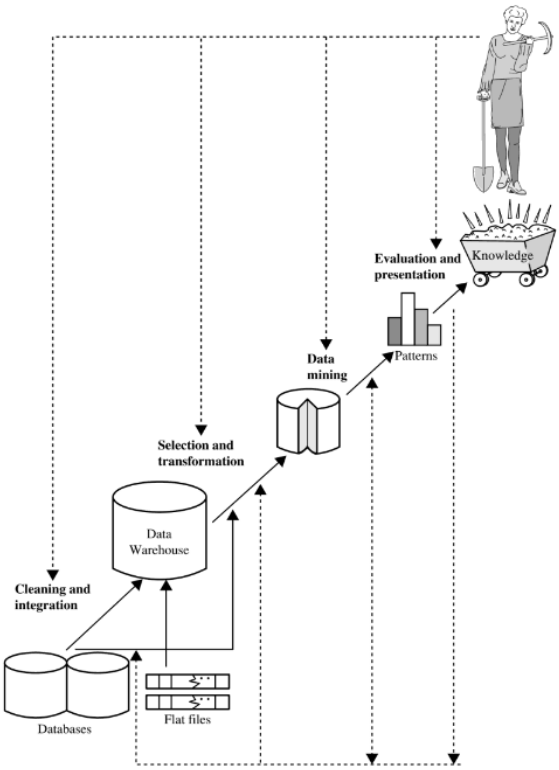
Data mining, juga dikenal sebagai penambangan data, adalah suatu proses mengekstraksi pola atau pengetahuan dari kumpulan data yang sangat besar [3]. Proses termasuk pengumpulan dan penggunaan data historis untuk menemukan pola, keteraturan, dan hubungan dalam set data yang sangat besar [12]. Dalam proses ini, teknik statistik dan algoritma digunakan untuk menemukan pola, tren, atau hubungan penting yang mungkin tidak terlihat dengan metode konvensional.

Akibatnya, penambangan data menjadi alat yang efektif untuk mengekstraksi pengetahuan yang memiliki potensi untuk mengubah cara kita memahami dan mengelola data.

Keseluruhan proses penemuan pengetahuan —mungkin karena lebih pendek daripada penemuan pengetahuan dari data, sering disebut *data mining* dalam industri, media, dan lingkungan penelitian. Oleh karena itu, kami mengadopsi pandangan luas tentang fungsionalitas penambangan data: *Data mining* adalah proses menemukan pola dan pengetahuan menarik dari kumpulan data yang sangat besar [5][13]. Data yang dialirkan secara dinamis ke dalam sistem, basis data, data gudang, web, atau repositori informasi lainnya adalah beberapa contoh sumber data [14].

Gambar menunjukkan proses penemuan pengetahuan 2.1 sebagai urutan berulang dari langkah-langkah berikut:

1. Pembersihan data dilakukan untuk mengurangi kebisingan serta memastikan konsistensi data.
2. Integrasi data, yaitu menggabungkan data dari berbagai sumber menjadi satu kesatuan.
3. Seleksi data, memilih data yang relevan dari basis data untuk keperluan analisis tertentu.
4. Transformasi data, melibatkan proses perubahan dan konsolidasi data ke dalam format yang sesuai untuk analisis, seperti melalui operasi agregasi atau ringkasan.
5. Penambangan data, merupakan proses utama yang menggunakan metode cerdas untuk menemukan pola dalam data.
6. Evaluasi pola, bertujuan untuk mengidentifikasi pola yang signifikan dan mewakili pengetahuan berdasarkan kriteria tertentu.
7. Penyajian pengetahuan, menggunakan teknik visualisasi dan representasi untuk menyampaikan hasil penambangan data kepada pengguna. [14].



Gambar 2.1 Proses *data mining* [13].

2.3 Klasifikasi

Klasifikasi adalah salah satu bentuk analisis data yang bertujuan untuk menghasilkan model yang dapat menggambarkan kelas-kelas data yang relevan. Model ini, yang dikenal sebagai pengklasifikasi, digunakan untuk memprediksi label kelas dalam bentuk kategoris (diskrit dan tidak berurutan) [13]. Beragam metode klasifikasi telah dikembangkan oleh para peneliti di bidang pembelajaran mesin, pengenalan pola, dan statistik. Sebagian besar algoritma awal dirancang untuk bekerja di memori dengan asumsi data berukuran kecil. Penelitian terkini dalam penambangan data memperluas pendekatan tersebut, menciptakan teknik klasifikasi dan prediksi yang dapat diskalakan untuk menangani data dalam jumlah besar yang tersimpan di memori [15].

Langkah pertama dalam proses klasifikasi adalah membangun pengklasifikasi untuk mendeskripsikan kelas-kelas data atau konsep yang telah ditentukan sebelumnya. Tahap ini dikenal sebagai fase pembelajaran (atau pelatihan), di mana algoritma klasifikasi mempelajari pola dengan menganalisis sekumpulan data

pelatihan yang terdiri dari tupel data beserta label kelas yang sesuai [14] . Tupel *X*, direpresentasikan oleh vektor atribut n-dimensi, *X = (x1, x2,..., xn)*, yang menggambarkan n pengukuran yang dilakukan pada tupel dari *n* atribut basis data, masing-masing, *A1, A2,..., An*. Setiap tupel, X, diasumsikan termasuk ke dalam salah satu kelas yang telah ditentukan sebelumnya, berdasarkan atribut tertentu yang dikenal sebagai atribut label kelas. Atribut ini bersifat diskrit, tidak berurutan, dan kategoris (atau nominal), karena setiap nilainya merepresentasikan sebuah kategori atau kelas tertentu. Tupel individu yang membentuk data pelatihan disebut tupel pelatihan, yang diambil secara acak dari basis data yang dianalisis. Dalam klasifikasi, tupel data sering disebut sebagai sampel, contoh, instansi, titik data, atau objek.

Karena setiap tupel pelatihan memiliki label kelas yang sudah ditentukan, proses klasifikasi juga dikenal sebagai *supervised learning*. Dalam pendekatan ini, pengklasifikasi "*supervised*" dengan informasi tentang kelas yang sesuai untuk setiap tupel pelatihan, sehingga memungkinkan model untuk belajar dengan cara terarah [14]. Pendekatan ini berbeda dengan *unsupervised learning* (atau klastering), di mana label kelas untuk setiap tupel pelatihan tidak tersedia. Selain itu, jumlah atau kelompok kelas yang akan dipelajari juga mungkin tidak diketahui sebelumnya, sehingga model harus menemukan struktur atau pola dalam data tanpa panduan eksplisit [16].

2.4 Machine Learning

Pembelajaran mesin (*machine learning*) adalah cabang dari ilmu komputer yang berfokus pada pengembangan algoritma dan model statistik untuk memungkinkan sistem komputer belajar dari data secara mandiri, tanpa perlu pemrograman eksplisit. Dengan kata lain, sistem ini dapat mengenali pola, memprediksi, dan membuat keputusan berdasarkan data yang diberikan, tanpa bergantung pada aturan yang telah ditetapkan sebelumnya [17].

Proses pembelajaran mesin melibatkan pemodelan data historis untuk melatih algoritma, sehingga model dapat menggeneralisasi pola yang ditemukan pada data baru. Algoritma pembelajaran mesin dapat dibagi menjadi beberapa kategori utama,

yaitu *supervised learning* (pembelajaran terawasi), *unsupervised learning* (pembelajaran tanpa pengawasan), dan *reinforcement learning* (pembelajaran penguatan). Setiap jenis memiliki karakteristik unik dan digunakan untuk tujuan serta penerapan yang berbeda sesuai dengan kebutuhan analisis data [18].

Penerapan pembelajaran mesin telah merambah ke berbagai bidang, termasuk:

- Analisis data: Mengidentifikasi tren, anomali, dan pola dalam data besar.
- Pengenalan pola: Mengidentifikasi objek dalam gambar, suara, atau teks.
- Prediksi: Memprediksi hasil peristiwa di masa depan, seperti harga saham atau cuaca.
- Rekomendasi sistem: Menyediakan rekomendasi yang dipersonalisasi, seperti rekomendasi produk atau konten.
- Otomasi: Mengotomatiskan tugas-tugas yang berulang, seperti pengenalan karakter optik atau klasifikasi dokumen.

Teknologi pembelajaran mesin telah menjadi pendorong utama dalam perkembangan kecerdasan buatan (*artificial intelligence*) dan telah mengubah lanskap berbagai industri. Dengan kemampuannya untuk mengolah data dalam skala besar dan kompleks, pembelajaran mesin membuka peluang baru untuk inovasi dan efisiensi.

a) Natural Language Proecess

Natural Language Processing (NLP) adalah bidang dalam kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. Tujuan utamanya adalah memungkinkan komputer untuk memahami, menganalisis, menafsirkan, dan menghasilkan bahasa alami secara efektif, sehingga dapat memberikan manfaat dalam berbagai aplikasi. Dalam NLP, berbagai teknik digunakan untuk memproses teks, seperti tokenisasi, *stemming*, *lemmatization*, *stopword removal*, dan *parsing* [3][13].

1. Tokenisasi adalah proses membagi teks menjadi bagian-bagian yang lebih kecil, seperti kata atau frasa, yang disebut token. Proses ini merupakan langkah awal dalam berbagai aplikasi NLP, seperti klasifikasi teks dan analisis sentimen.

2. *Stopword Removal* mengacu pada penghapusan kata-kata umum yang tidak memberikan banyak informasi dalam konteks analisis, seperti "dan", "dari", atau "dengan". Penghapusan stopword membantu mengurangi kompleksitas data tanpa kehilangan makna penting.
3. *Stemming* dan *lemmatization* adalah teknik yang digunakan untuk mengubah kata-kata ke bentuk dasarnya. Stemming menghilangkan akhiran kata secara sederhana untuk mendapatkan bentuk dasar, sementara lemmatization mempertimbangkan konteks linguistik dan mengubah kata ke bentuk leksikal yang lebih tepat atau umum.
4. *Named Entity Recognition* (NER) adalah teknik NLP yang digunakan untuk mengidentifikasi entitas penting dalam teks, seperti nama orang, tempat, organisasi, dan tanggal. Teknik ini banyak digunakan dalam aplikasi seperti ekstraksi informasi dan pencarian berbasis teks.
5. *Sentiment Analysis* adalah salah satu aplikasi populer NLP yang bertujuan untuk mengidentifikasi sikap atau perasaan dalam teks, apakah positif, negatif, atau netral. Analisis sentimen banyak digunakan dalam analisis opini, ulasan pelanggan, dan media sosial [19][20].
6. *Word Embeddings* seperti Word2Vec atau GloVe memungkinkan komputer untuk memahami makna kata dalam konteks, dengan merepresentasikan kata-kata dalam bentuk vektor. Ini memungkinkan model NLP untuk menangkap hubungan semantik antar kata dan digunakan dalam tugas-tugas seperti klasifikasi teks, penerjemahan mesin, dan *chatbot*.

Dengan berbagai teknik NLP ini, banyak aplikasi praktis yang dapat diterapkan, seperti analisis data teks dalam jumlah besar, *chatbots*, mesin pencari, dan terjemahan otomatis. Tantangan dalam NLP adalah menangani kompleksitas dan ambiguitas bahasa manusia, yang seringkali penuh dengan sinonim, homonim, dan nuansa konteks [21].

SpaCy adalah pustaka sumber terbuka yang dirancang untuk Pemrosesan Bahasa Alami (NLP) tingkat lanjut menggunakan Python. SpaCy difokuskan pada penggunaan dalam produksi dan membantu membangun aplikasi yang dapat memproses serta "memahami" teks dalam jumlah besar. Pustaka ini dapat

digunakan untuk mengembangkan sistem ekstraksi informasi, pemahaman bahasa alami, atau melakukan praproses teks untuk pembelajaran mendalam.

b) K-Nearest Neighbor

Algoritma K-Nearest Neighbors (KNN) adalah metode *supervised learning* yang digunakan untuk masalah klasifikasi dan regresi. Algoritma ini pertama kali dikembangkan oleh Evelyn Fix dan Joseph Hodges pada tahun 1951, kemudian diperkenalkan dan diperluas oleh Thomas Cover. KNN bekerja dengan mengklasifikasikan data berdasarkan kedekatannya dengan data lain yang sudah diketahui labelnya [5].

Algoritma K-NN adalah algoritma pembelajaran mesin yang serbaguna dan banyak digunakan yang terutama digunakan karena kesederhanaan dan kemudahan implementasinya. Algoritma ini tidak memerlukan asumsi apa pun tentang distribusi data yang mendasarinya. Algoritma ini juga dapat menangani data numerik dan kategorikal, menjadikannya pilihan yang fleksibel untuk berbagai jenis kumpulan data dalam tugas klasifikasi dan regresi. Algoritma ini adalah metode nonparametrik yang membuat prediksi berdasarkan kesamaan titik data dalam kumpulan data tertentu. K-NN kurang sensitif terhadap outlier dibandingkan dengan algoritma lainnya.

Algoritma K-Nearest Neighbors (K-NN) berfungsi dengan mencari K tetangga terdekat dari titik data tertentu menggunakan metrik jarak, seperti jarak Euclidean. Kelas atau nilai dari titik data tersebut kemudian ditentukan berdasarkan suara mayoritas (untuk klasifikasi) atau rata-rata (untuk regresi) dari K tetangga terdekat. Pendekatan ini memungkinkan algoritma untuk menyesuaikan diri dengan pola yang bervariasi dan membuat prediksi berdasarkan struktur lokal data [6].

Seperti yang kita ketahui bahwa algoritma KNN membantu kita mengidentifikasi titik atau grup terdekat untuk suatu titik kueri. Namun untuk menentukan grup terdekat atau titik terdekat untuk suatu titik kueri, kita memerlukan beberapa metrik. Untuk tujuan ini, kita menggunakan metrik jarak berikut:

Euclidean Distance

$$d(x, X_i) = \sqrt{\sum_{j=1}^{d} \left(x_j - X_{ij}\right)^2}$$

Manhattan Distance

$$d(\mathrm{x}, \mathrm{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

Minkowski Distance

$$d(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^p\right)^{\frac{1}{p}}$$

Dari rumus di atas kita dapat mengatakan bahwa ketika p = 2 maka rumus tersebut sama dengan rumus untuk jarak Euclidean dan ketika p = 1 maka kita memperoleh rumus untuk jarak Manhattan.

Metrik yang dibahas di atas adalah yang paling umum digunakan saat menangani masalah *Machine Learning*, tetapi ada juga metrik jarak lain seperti *Hamming Distance* yang berguna saat menangani masalah yang memerlukan perbandingan yang tumpang tindih antara dua vektor yang isinya dapat berupa nilai *boolean* maupun *string*.

c) Naïve Bayes

Klasifikasi Naïve Bayes atau Naïve Bayes Classifier adalah sekumpulan algoritma klasifikasi yang didasarkan pada Teorema Bayes. Meskipun mengasumsikan independensi antar fitur (yang dianggap "naive"), metode ini sangat populer karena kesederhanaannya dan efisiensinya dalam penerapan pembelajaran mesin, terutama dalam tugas klasifikasi teks seperti analisis sentimen dan filter spam. Klasifikasi Naïve Bayes bekerja dengan menghitung probabilitas posterior untuk setiap kelas berdasarkan probabilitas fitur, lalu memilih kelas dengan probabilitas tertinggi [13].

Algoritma Naïve Bayes digunakan untuk masalah klasifikasi. Algoritma ini banyak digunakan dalam klasifikasi teks. Dalam tugas klasifikasi teks, data mengandung

dimensi tinggi (karena setiap kata mewakili satu fitur dalam data). Algoritma ini digunakan dalam penyaringan spam, deteksi sentimen, klasifikasi peringkat, dll. Keuntungan menggunakan Naïve Bayes adalah kecepatannya. Cepat dan mudah membuat prediksi dengan dimensi data yang tinggi.

Model ini memprediksi probabilitas sebuah instance termasuk dalam kelas dengan serangkaian nilai fitur tertentu. Model ini merupakan pengklasifikasi probabilistik. Hal ini karena model ini mengasumsikan bahwa satu fitur dalam model tidak bergantung pada keberadaan fitur lain. Dengan kata lain, setiap fitur berkontribusi pada prediksi tanpa hubungan satu sama lain. Dalam dunia nyata, kondisi ini jarang terpenuhi. Model ini menggunakan teorema Bayes dalam algoritma untuk pelatihan dan prediksi [16].

Teorema Bayes digunakan untuk menghitung probabilitas terjadinya suatu peristiwa berdasarkan informasi yang sudah diketahui tentang peristiwa lain. Teorema ini menyatakan probabilitas posterior dari suatu peristiwa berdasarkan probabilitas prior dan likelihood.

Secara matematis, Teorema Bayes dapat dinyatakan dengan persamaan berikut:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

di mana A dan B adalah kejadian dan P(B) ≠ 0

- **P(A)**: Ini adalah **probabilitas apriori** dari peristiwa A, yaitu probabilitas kejadian A sebelum adanya bukti (kejadian B). Ini menggambarkan pengetahuan awal tentang A sebelum bukti baru (B) muncul.
- **P(B)**: Ini adalah **probabilitas marjinal** dari bukti B, yaitu probabilitas terjadinya B tanpa mempertimbangkan A. Dalam konteks ini, B sering disebut sebagai "bukti" yang digunakan untuk memperbarui keyakinan tentang A.
- **P(A|B)**: Ini adalah **probabilitas posteriori** dari A, yaitu probabilitas terjadinya A setelah bukti B terlihat. Ini adalah nilai yang ingin kita hitung, yakni probabilitas A terjadi dengan mempertimbangkan bukti B yang sudah ada.

- **P(B|A)**: Ini adalah **probabilitas likelihood** atau kemungkinan bahwa bukti B akan terwujud jika A benar. Dalam konteks pembelajaran mesin, ini menggambarkan seberapa besar kemungkinan kita mengamati bukti B ketika hipotesis A benar.

Secara singkat, dalam Teorema Bayes, kita mencoba untuk menghitung probabilitas kejadian A (probabilitas posterior) setelah kita memiliki bukti B (atau data terkait), yang memperbarui keyakinan kita tentang A [5].

Aplikasi teorema Bayes pada kumpulan data sebagai berikut:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

di mana, y adalah variabel kelas dan X adalah vektor fitur dependen (berukuran n) di mana:

$$X = (x_1, x_2, x_3, \dots, xn)$$

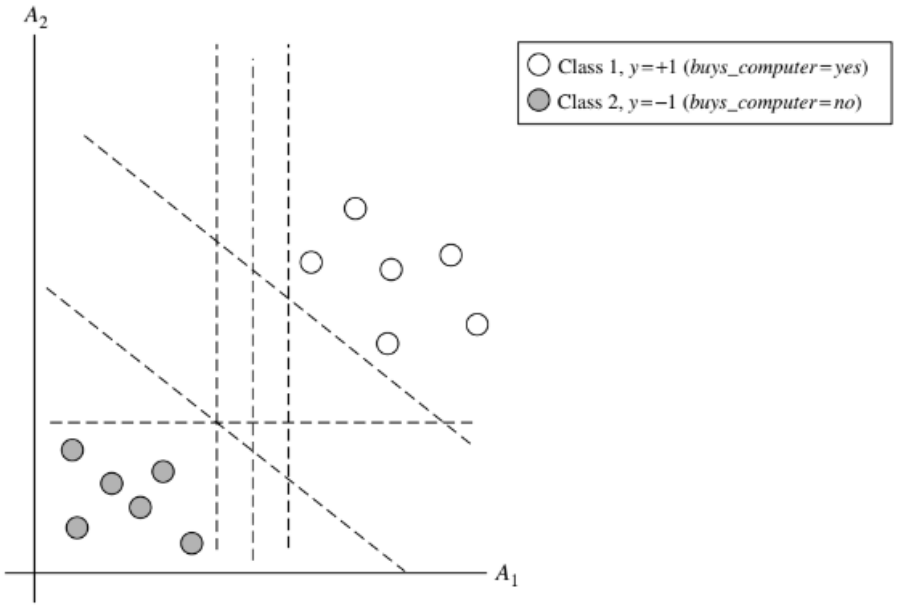
d) Support Vector Machine

Support Vector Machine (SVM) adalah metode yang digunakan untuk klasifikasi data baik yang linear maupun non-linear. SVM bekerja dengan melakukan pemetaan non-linear untuk mengubah data pelatihan ke dalam dimensi yang lebih tinggi. Dalam dimensi baru ini, SVM mencari hyperplane yang memisahkan data dari satu kelas ke kelas lainnya secara optimal. *Hyperplane* ini berfungsi sebagai "batas keputusan" yang memisahkan dua kelas.

Dengan pemetaan non-linear yang tepat ke dimensi yang cukup tinggi, data dari dua kelas akan selalu dapat dipisahkan oleh *hyperplane*. SVM menentukan *hyperplane* ini dengan menggunakan *support vectors*, yaitu tupel pelatihan yang dianggap "esensial" atau penting, dan margin, yang didefinisikan oleh *support vectors* dan menggambarkan jarak antara *hyperplane* dan titik data terdekat dari kedua kelas. SVM berusaha untuk memaksimalkan margin ini agar memperoleh pemisahan yang optimal [13][16].

Meskipun waktu pelatihan, bahkan untuk SVM tercepat, bisa sangat lambat, algoritma ini sangat akurat karena kemampuannya dalam memodelkan batas keputusan non-linier yang kompleks. SVM cenderung lebih tahan terhadap *overfitting* dibandingkan metode lainnya. Selain itu, vektor pendukung yang ditemukan memberikan gambaran singkat tentang model yang telah dipelajari. SVM dapat digunakan untuk prediksi numerik serta klasifikasi. Metode ini telah diterapkan di berbagai bidang, seperti pengenalan digit tulisan tangan, pengenalan objek, identifikasi pembicara, dan prediksi deret waktu untuk uji tolok ukur.

Contoh kasus yang paling sederhana gambar 2.2 —masalah dua kelas di mana kelas-kelas tersebut dapat dipisahkan secara linier. Misalkan himpunan data *D* diberikan sebagai *(X1, y1), (X2, y2), ... , (X|D|, y|D|)*, di mana *Xi* adalah himpunan tupel pelatihan dengan label kelas terkait, yi. Setiap yi dapat mengambil salah satu dari dua nilai, baik +1 atau −1 (yaitu, *yi ∈{+1, − 1})*,



Gambar 2.2 Kasus *buys computer support vector machine* 2 dimensi [13].

sesuai dengan kelas *buys computer = yes* dan *buys computer = no*, berturut-turut. Untuk membantu visualisasi, mari kita pertimbangkan contoh berdasarkan dua atribut input, *A1* dan *A2*. Dari grafik, kita melihat bahwa data *2-D* dapat dipisahkan secara linear (atau "linear," singkatnya), karena garis lurus dapat ditarik untuk memisahkan semua tupel kelas +1 dari semua tupel kelas −1 [13].

2.6 VS Code

VS Code dikenal *Visual Studio Code,* merupakan salah satu *code editor* populer bagi pengembang perangkat lunak. Ini cepat, ringan, dapat disesuaikan, dan berisi dukungan bawaan untuk JavaScript, TypeScript, dan ekstensi Node.js untuk bahasa lain, termasuk C++*,* Python, dan PHP. Fitur-fitur seperti kemampuan *debugging*, kontrol Git yang tertanam, penyorotan sintaksis, cuplikan kode, dan dukungan penyelesaian kode cerdas IntelliSense—yang beberapa di antaranya membedakannya dari pesaing—membantu menjadikan *Visual Studio Code* sebagai solusi yang mengesankan dan unik.

Visual Studio Code adalah alat pengeditan dan *Debugging End-to-End* untuk pengembangan perangkat lunak yang membantu pengembang menjadi mudah dan lebih produktif [22]. Panduan teknis mencakup semua komponen penting perangkat lunak, termasuk fitur pengeditan ruang kerja, fungsionalitas tingkat lanjut seperti pemfaktoran ulang kode dan pengikatan kunci, serta integrasi dengan Grunt, Gulp*,* NPM, dan alat eksternal lainnya. Pengguna baru, pengembang berpengalaman, dan mereka yang mempertimbangkan untuk pindah dari alat pengembang lain akan mendapatkan manfaat dari informasi terperinci namun mudah diikuti tentang *Visual Studio Code* [23].

2.7 Python

Python adalah bahasa pemrograman dilengkapi perangkat lunak *interpreter* Python yang membaca kode sumber (ditulis dalam bahasa Python) dan menjalankan instruksinya [24][25]. Nama Python berasal dari grup komedi surealis Inggris Monty Python, bukan dari ular [26]. Programmer Python sering disebut Pythonistas, dan referensi Monty Python dan ular biasanya menghiasi tutorial dan dokumentasi Python.

Bahasa pemrograman Python menawarkan berbagai konstruksi sintaksis, fungsi pustaka standar, dan fitur dalam lingkungan pengembangan interaktif. Python memungkinkan pengguna untuk mengetik langsung ke dalam shell interaktif, yang juga dikenal sebagai REPL (*Read-Evaluate-Print Loop*). Fitur ini memungkinkan eksekusi instruksi Python secara langsung satu per satu dan memberikan hasil

secara instan. Menggunakan shell interaktif sangat bermanfaat untuk mempelajari cara kerja instruksi dasar Python [26].

2.8 Looker Studio

Google Looker Studio (sebelumnya Data Studio) adalah alat yang digunakan untuk memvisualisasikan data [27]. Google telah mengganti nama Data Studio menjadi Looker Studio sejak Google mengakuisisi alat visualisasi Looker. Google telah menyatukan produk kecerdasan bisnis mereka di bawah payung Looker dengan menggabungkan fungsionalitas Data Studio, Looker, dan kecerdasan buatan dalam Looker Studio. Semua fungsionalitas dari Data Studio akan tetap sama di Looker Studio.

Looker Studio adalah alat berbasis *cloud* yang berarti Anda dapat mengaksesnya dari perangkat/browser apa pun selama memiliki koneksi internet yang stabil. Looker Studio merupakan alat visualisasi gratis yang memungkinkan untuk membuat *dashboard* dan laporan. Google juga menyediakan Looker Studio Pro, yang merupakan versi berbayar dari Looker Studio [28].

Looker Studio adalah alat visualisasi yang hebat karena serbaguna dan mudah untuk membuat, berbagi, dan berkolaborasi tanpa biaya. Berikut adalah beberapa manfaat utama menggunakan Looker Studio sebagai alat visualisasi [29].

1. Sumber data yang beragam: dapat menghubungkan Looker Studio ke berbagai sumber data dan mengumpulkan serta menggabungkan data dalam satu laporan. Dengan cara ini, dapat mengukur aktivitas pemasaran di berbagai platform dan saluran serta menghasilkan wawasan lintas platform dan multi-saluran.
2. Kustomisasi tanpa batas: dapat sepenuhnya menyesuaikan laporan Looker Studio. Dapat menggunakan kanvas kosong untuk mendesain laporan sendiri dari awal, atau dapat menggunakan template Looker Studio. Ini membuat Looker Studio menjadi cara yang sangat nyaman dan bebas masalah untuk mendesain laporan dan dashboard untuk berbagai platform data seperti Google Analytics, Google Ads, iklan YouTube, dll.

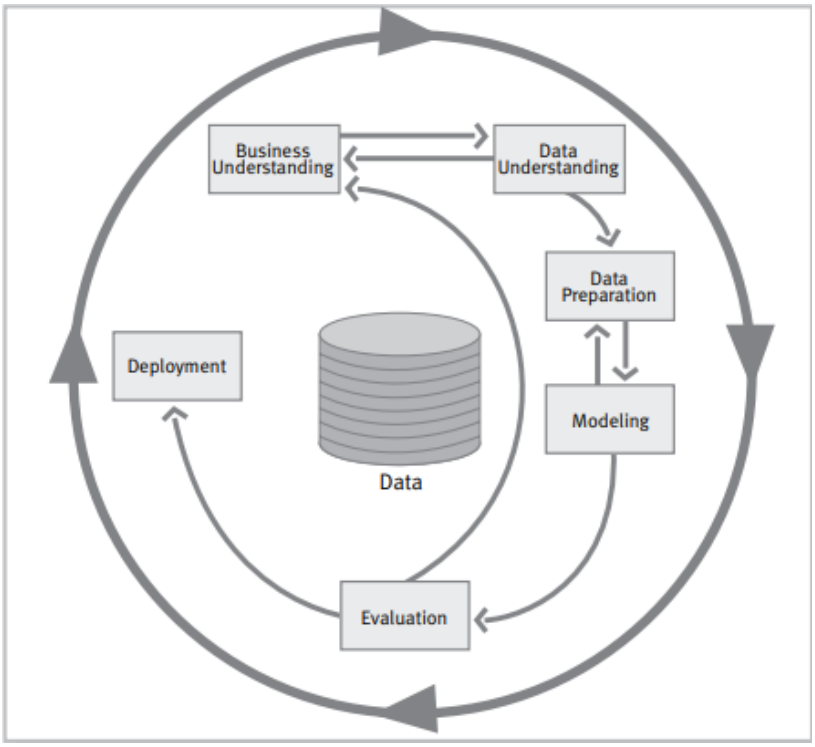
3. Pelaporan dinamis dan real-time: Laporan Looker Studio sangat dinamis. Ini berarti dapat menerapkan berbagai jenis filter pada sumber data dan mempersempit data berdasarkan rentang tanggal, waktu, pengguna, kategori perangkat, negara, dll. Selain itu, laporan-laporan ini dapat dibuat secara *real-time* dengan menarik data secara real-time saat tersedia di sumber data terkait.
4. Berbagi Laporan dan Kolaborasi: Laporan Looker Studio dapat dibagikan dalam berbagai format dengan menjadwalkan pengiriman email, membuat tautan untuk laporan, atau mengunduh laporan sebagai PDF. Dapat menambahkan rekan kerja ke laporan untuk tujuan pengeditan dan membaca, membuat kolaborasi mudah di seluruh tim.
5. Tanpa biaya: Salah satu keunggulan besar menggunakan Looker Studio adalah sepenuhnya gratis untuk digunakan. Selain itu, tidak ada batasan jumlah pengguna per akun Looker Studio, yang membuatnya lebih bermakna di organisasi besar [29].

2.9 CRISP-DM

Cross Industry Standard Process for Data-Mining (CRISP-DM) adalah metode yang digunakan dalam industri untuk menyelesaikan berbagai masalah bisnis yang berkaitan dengan *data mining* [10][30]. Model CRISP-DM menggambarkan siklus proyek data mining, yang terdiri dari enam tahapan utama, yaitu pemahaman bisnis (*business understanding*), pemahaman data (*data understanding*), persiapan data (*data preparation*), pemodelan (*modeling*), evaluasi (*evaluation*), dan terakhir penerapan (*deployment*). [31].

Tahapan awal pada alur model CRISP-DM yaitu tahapan *business understanding*, berdasarkan data yang dimiliki akan dilihat perspektif dari segi bisnis untuk mencapai tujuan bisnis. Kedua, tahapan *data understanding*, pada tahapan ini data-data dikumpulkan dan dipahami, pada tahapan ini juga dapat kembali ke tahapan sebelumnya yaitu *business understanding* untuk memastikan bahwa data-data yang dikumpulkan dapat digunakan untuk mencapai tujuan. *Data preparation*, pada tahapan ini data-data yang telah dikumpulkan dilakukan persiapan data seperti

menyeleksi, mengintegrasikan, membersihkan dan memformat data. Kemudian data yang sudah disesuaikan masuk tahapan *modeling*, pada tahapan ini melakukan pemilihan teknik model *data mining* yang ingin diterapkan, pada tahapan ini juga dapat kembali ke tahapan sebelumnya yaitu tahapan *data preparation*, hal tersebut dikarenakan jika dalam tahapan *modeling* ditemukan data-data yang tidak diinginkan dan mengganggu hasil pemodelan maka dapat kembali ke tahapan *data preparation* untuk memperbaiki permasalah data [5]. Setelah tahapan *modelling* selesai, data dipersiapkan untuk tahapan *evaluation*, pada tahapan ini akan melakukan evaluasi model yang telah dibangun untuk memastikan bahwa tujuan bisnis telah terjawab di tahapan *business understanding* dan telah memiliki nilai performa model yang baik. Pada tahapan *evaluation* terdapat siklus balik kembali ke *business understanding* dikarenakan tidak semua kasus harus dilakukan *deployment*, terdapat kasus jika sudah mendapatkan informasi pada tahapan *evaluation* akan melakukan iterasi kembali di tahapan *business understanding* [10]. Tahapan-tahapan tersebut dikembangkan dalam bentuk diagram CRISP-DM pada gambar 2.3.



Gambar 2.3 Tahapan-tahapan CRISP-DM [10].

2.10 Penelitian Terkait

Dalam penyusunan skripsi ini, penulis banyak mendapatkan inspirasi dari penelitian-penelitian sebelumnya. Adapun penelitian yang berkaitan mengenai teknik pemrosesan teks, penelitian *data mining*, algoritma klasifikasi SVM dan metode penelitian CRISP-DM antara lain:

Tabel 2.1 Penelitian Terkait

No	Deskripsi
1	**Judul:** Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)
	Penulis: Primandani Arsi, Retno Waluyo
	Tahun: 2021
	Metode: *data crawling* dan algoritma SVM
	Data: Penelitian ini menggunakan data tweet yang berkaitan dengan topik pemindahan ibu kota Indonesia, yang diperoleh melalui metode *crawling* pada API Twitter. Dalam penelitian ini, diusulkan penggunaan metode Support Vector Machine (SVM) untuk diterapkan pada tweet-tweet mengenai topik pemindahan ibu kota Indonesia dengan tujuan melakukan klasifikasi sentimen pada media sosial Twitter. Proses klasifikasi dilakukan dengan membagi tweet menjadi dua kelas, yaitu positif dan negatif [19].
2	**Judul:** *A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification*
	Penulis: Kanish Shah, Hanil Patel, Devanshi Sanghvi, Manan Shah
	Tahun: 2020
	Metode: Random Forest, KNN, dan *evaluation methode*
	Data: Sistem klasifikasi teks dibagi menjadi empat bagian yaitu pra-pengolahan teks, representasi teks, implementasi pengklasifikasi dan klasifikasi. Pada bagian implementasi pemodelan, penulis secara terpisah memilih dan membandingkan regresi logistik, random forest, dan KNN sebagai algoritma klasifikasi. Penulis memutuskan untuk menunjukkan perbandingan berdasarkan lima parameter yaitu *presisi, akurasi, F1-score, support,* dan *confusion matrix* [32].

Tabel 2.1 Penelitian Terkait (1)

No	Deskripsi
3	**Judul:** Implementasi *Text-Mining* untuk Analisis Sentimen pada Twitter dengan Algoritma Support Vector Machine
	Penulis: Aditiya Hermawan, Indrico Jowensen, Junaedi, Edy
	Tahun: 2023
	Metode: *data preparation*, *data mining*, *evaluation methode*, dan algoritma SVM
	Data: Metode yang digunakan dalam penelitian ini adalah eksperimen, di mana data diuji secara langsung dengan pendekatan *Text Mining* yang terdiri dari lima proses utama, yaitu *pre-processing* teks, transformasi teks, seleksi fitur, *data mining*, dan evaluasi. Model yang dihasilkan untuk melakukan klasifikasi sentimen menggunakan SVM diterapkan secara langsung untuk mengklasifikasikan sentimen dari suatu topik di Twitter. Proses ini dilakukan dengan mencari kata-kata yang relevan untuk diklasifikasikan, dan kemudian memperoleh informasi berupa klasifikasi sentimen dari kalimat-kalimat yang terkait [33].
4	**Judul:** *Data Mining for Predicting the Amount of Coffee Production Using CRISP-DM Method*
	Penulis: Ali Khumaidi
	Tahun: 2020
	Metode: CRISP-DM
	Data: Metode yang digunakan dalam penelitian ini adalah CRISP-DM dan algoritma regresi linier berganda untuk memprediksi jumlah kopi yang diproduksi serta mengetahui hubungan antar variabel. Langkah-langkah yang dilakukan meliputi pemahaman bisnis, pemahaman data, penyiapan data, pemodelan, dan evaluasi. Dataset yang digunakan terdiri dari 170 data, yang setelah melalui tahap persiapan data, disusun menjadi 150 data dengan 5 atribut yang tercantum dalam tabel [34].

Tabel 2.1 Penelitian Terkait (2)

No	Deskripsi
5	**Judul:** Analisis Sentimen Wisatawan terhadap Kualitas Layanan Hotel dan Resort di Lombok Menggunakan SERVQUAL dan CRISP-DM
	Penulis: Yerik Afrianto Singgalen
	Tahun: 2023
	Metode: SERVQUAL, CRISP-DM, Naïve Bayes, dan SVM
	Data: Tujuan dari penelitian ini adalah untuk mengevaluasi kualitas layanan hotel dengan menggunakan kerangka kerja *Quality of Service* (SERVQUAL). Hasil klasifikasi data sentimen tamu hotel didasarkan pada algoritma Naive Bayes Classifier (NBC) dan Support Vector Machine (SVM), yang dilakukan sesuai dengan tahapan *Standard Process for Data Mining Cross-Industry* (CRISP-DM). Kerangka kerja CRISP-DM terdiri dari enam tahap, yaitu pemahaman bisnis, pemahaman data, pembuatan data, modeling, evaluasi, dan penyebaran. [20].
6	**Judul:** Analisis Sentimen Pada Ulasan Pengguna Aplikasi Bibit Dan Bareksa Dengan Algoritma KNN
	Penulis: Aluisius Dwiki Adhi Putra, Safitri Juanita
	Tahun: 2021
	Metode: CRISP-DM, KNN, *accuracy, precision, recall, f1-score,* dan *confusion matrix*
	Data: Dengan menganalisis ulasan pengguna tentang aplikasi bibit dan bareksa, penelitian ini bertujuan untuk memberi masyarakat saran tentang aplikasi investasi online mana yang lebih baik dan aman. Untuk penelitian ini, model CRISP-DM (Cross Industry Standard Process for Data Mining) digunakan. Algoritma yang digunakan dalam penelitian ini adalah KNN. Hasil penelitian dinilai dengan menggunakan nilai akurasi dan recall yang dihasilkan dari tahapan modeling menggunakan algoritma KNN dan perbandingan 60:40. [35].

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian dilakukan di kampus Universitas Lampung selama 8 bulan yang dimulai dari bulan Januari s.d. Agustus 2024. Rincian jadwal kegiatan penelitian terdapat pada tabel 3.1.

Tabel 3.1 Jadwal Penelitian

No	Aktivitas	2024							
		Jan	Feb	Mar	Apr	Mei	Jun	Jul	Ags
1	Studi Literatur								
2	*Business Understanding*								
3	*Data Understanding*								
4	*Data Preparation*								
5	*Modelling*								
6	*Evaluation*								
7	*Deployment*								
8	Pelaporan								

3.2 Alat & Bahan Penelitian

Dalam penelitian ini, dibutuhkan alat dan bahan untuk mendukung kelancaran proses penelitian. Alat dan bahan tersebut antara lain:

3.2.1 Alat

Penelitian dan penyusunan tugas akhir ini menggunakan alat perangkat keras dan perangkat lunak dengan spesifikasi sebagai berikut:

1. Perangkat Keras

   Perangkat keras penunjang penelitian ini, sebagai berikut:

   a. Sebuah *Personal Computer (PC)* dengan spesifikasi:

      - Prosesor : *Intel(R) Core* i7-12700H 2.30 *GHz*
      - RAM : RAM 16 GB
      - Memori : 512 *GB*
      - Sistem Operasi : *Windows 11 64-bit*

2. Perangkat Lunak

   Perangkat lunak penunjang penelitian ini, sebagai berikut:

   a. VS Code

   VS Code digunakan untuk melakukan pengkodean *Python* dalam melakukan penarikan data dari *database* dengan *data scrapping* untuk dilakukan analisis serta *data preparation* seperti menghilangkan data yang tidak digunakan, dan memeriksa kesalahan pada data yang lainnya VS Code juga digunakan dalam melakukan analisis *data mining* menggunakan metode dan algoritma yang diinginkan. Serta digunakan untuk memvisualisasikan data berupa gambar yang informatif agar lebih mudah dipahami dan dibagikan kepada *stakeholder* terkait.

   b. Looker Studio

   Looker Studio digunakan pada tahap penyebaran, di mana seluruh informasi hasil tahap pemodelan divisualisasikan dalam bentuk yang disesuaikan. Bentuk visualisasi berbentuk tabel untuk memberi gambaran visibilitas data secara menyeluruh, dan bentuk diagram untuk mempermudah dalam pengambilan kesimpulan.

   c. Jobstreet

Seluruh data diambil dari API Jobstreet Indonesia. Set data berisi sub-kategori lowongan pekerjaan bidang teknologi dan informasi, dengan atribut pendukung lain.

3.2.2 Bahan Penelitian

Bahan yang diperlukan dalam penelitian ini merupakan data lowongan kerja bidang teknologi dan informasi yang diambil dari API Jobstreet Indonesia (*www.jobstreet.com*). Pengambilan data dengan teknik *data scrapping* dengan bahasa Python di VS Code. Periode pengambilan data pada tanggal Januari, April, dan Juni 2024. Data berupa file *excel* dengan format csv sebanyak 2340 baris.

3.3 Metode Penelitian

Metode dalam penelitian ini menggunakan CRISP-DM. Berikut merupakan alur tahapan metode CRISP-DM:

3.3.1 Studi Literatur

Tahapan studi literatur dilakukan untuk mempelajari ilmu dan penelitian terkait, studi literatur diambil dari penelitian sebelumnya yang bersumber dari buku, jurnal, artikel dan prosiding. Referensi ilmu pada studi literatur berguna untuk mendukung penelitian ini, seperti pemrosesan teks, *data mining*, algoritma *machine learning*, pustaka Python dan metode pengembangan CRISP-DM.

Pustaka Python, juga dikenal sebagai *library* Python, adalah kumpulan kode yang telah dibuat sebelumnya dan dapat digunakan kembali dalam program Python. Bayangkan sebuah lemari dengan berbagai alat. Ketika ingin menyelesaikan tugas tertentu, cukup menggunakan alat yang sudah ada di lemari itu daripada membuatnya dari awal [14].

Berikut pustaka Python yang digunakan dalam proses penelitian ini [5][21]:

NLTK (*Natural Language Toolkit*): pustaka ini merupakan salah satu yang paling lengkap dan populer. NLTK menyediakan berbagai alat untuk tokenisasi, *stemming, tagging part-of-speech, parsing*, dan banyak lagi.

spaCy: dikenal karena kecepatan dan efisiensi, spaCy sangat cocok untuk tugas-tugas yang membutuhkan pemrosesan teks dalam skala besar.

Gensim: spesialis dalam topik *modeling, document similarity*, dan *word embeddings*. Gensim sangat berguna untuk analisis teks yang lebih kompleks.

TextBlob: pustaka yang dibangun di atas NLTK dan *pattern*, TextBlob menyediakan antarmuka yang lebih sederhana untuk tugas-tugas umum seperti *sentiment analysis* dan *part-of-speech tagging*.

scikit-learn: meskipun bukan khusus untuk pemrosesan teks, scikit-learn menyediakan berbagai algoritma *machine learning* yang sangat berguna untuk tugas-tugas seperti klasifikasi teks dan *clustering*.

pandas: pustaka ini sangat berguna untuk manipulasi data, termasuk data teks. Pandas dapat digunakan untuk membersihkan data, melakukan transformasi, dan menganalisis data.

NumPy: pustaka fundamental untuk komputasi ilmiah dalam Python. Ini menyediakan dukungan untuk array multidimensi, matriks, dan berbagai fungsi matematika tingkat tinggi.

Matplotlib: pustaka untuk membuat visualisasi data dalam Python.

Pyplot: bagian dari Matplotlib yang menyediakan antarmuka mirip MATLAB untuk membuat plot.

Counter: sebuah kelas yang digunakan untuk menghitung frekuensi elemen dalam suatu *iterable* (seperti *list, tuple*, atau *string*).

WordCloud: sebuah kelas yang digunakan untuk menghasilkan visualisasi kata-kata dalam bentuk awan kata (*word cloud*).

re adalah singkatan dari *regular expression* atau ekspresi reguler. Ekspresi reguler adalah suatu pola pencarian yang digunakan untuk mencocokkan karakter atau kumpulan karakter dalam sebuah teks.

Seaborn untuk membuat visualisasi data. Memberikan tampilan yang lebih menarik dan informatif dibandingkan dengan matplotlib dasar. Seaborn berguna

untuk mengeksplorasi data dan menemukan pola-pola yang mungkin tidak terlihat dengan jelas dalam data mentah.

Metode penelitian CRISP-DM berjalan dengan membagi proses menjadi beberapa tahap, seperti *business understanding* (tahap pemahaman bisnis), *data understanding* (pemahaman data), *data preparation* (persiapan data), *modelling* (tahap pemodelan), *evaluation* (tahap evaluasi) dan *deployment* (tahap penyebaran) [10].

3.3.2 *Business Understanding* (Tahap Pemahaman Bisnis)

Dalam tahap pemahaman bisnis bertujuan untuk menentukan urgensi dilakukan penelitian, menilai situasi, menentukan tujuan *data mining*, dan membuat rencana proyek. Tahapan ini memiliki pokok bahasan memahami rumusan masalah, proses mencapai tujuan penelitian, proses mendapatkan data pendukung, potensi pemanfaatan data dan analisis untuk memperoleh model terbaik.

3.3.3 *Data Understanding* (Tahap Pemahaman Data)

Serangkaian tindakan dilakukan untuk mendapatkan pemahaman yang lebih baik tentang data yang digunakan dalam penelitian ini pada tahap pemahaman data, di mana fokus utamanya adalah pengumpulan data, eksplorasi data, verifikasi kualitas data, visualisi data dan *exploratory data analysis* (EDA) [36]. Pada tahapan ini dilakukan proses pengumpulan data awal, mengidentifikasi kualitas data yang ada, mendeteksi pengetahuan dari data, kemudian melakukan analisia data untuk membentuk hipotesis dari informasi yang tersembunyi. Semua tindakan ini sangat penting untuk memahami karakteristik data, menemukan potensi masalah, dan mendapatkan pemahaman awal tentang situasi sebelum melanjutkan ke tahap berikutnya.

Proses pengumpulan data menggunakan teknik *data scraping*, data diambil pada API Jobstreet Indonesia. Varibel yang berhasil diambil seperti, *job titles, company, descriptions, location, sub categorys, job types, salarys,* dan *date ingestions*. Variabel terkumpul pada sebuah data dengan format CSV *(comma separated values)*.

Pemahaman data awal merupakan eksplorasi data mencakup menghitung jumlah record data, melihat jumlah *record* data berdasarkan variabel, *output* di atas menampilkan variabel pendukung, disertai dengan jumlah *record, unique words, top row,* dan *freq*. Verifikasi kualitas data berguna untuk melihat tingkat kualitas data yang tersedia sehingga sumber daya data dapat diberdayakan secara maksimal, seperti memeriksa *missing value*, pemeriksaan distribusi panjang teks pada variabel *descriptions*. Visualisasi data berguna untuk melihat perhitungan statistik jumlah sebaran data dalam angka, diagram batang dan *wordcloud. Exploratory Data Analysis* (EDA) menampilkan n-gram seperti unigram, bi-gram, dan tri-gram. Dan scatter plot untuk memeriksa hubungan varibel *subcategory* dengan *descriptions*.

3.3.4 *Data Preparation* (Tahap Persiapan Data)

Tahap persiapan data bertujuan menyeleksi data yang akan digunakan, mengintegrasikan data terpilih, membersihkan data yang tidak terpakai, dan memformat data agar dapat diolah dengan menggunakan teknik *data mining*. Pustaka yang digunakan adalah regex dan spacy. Regex untuk normalisasi dan pembersihan data, sedangkan spacy untuk transformasi data. Tahap persiapan merupakan proses yang dilakukan untuk menyiapkan data dengan menyesuaikan *dataset* agar sesuai dengan kebutuhan yang akan digunakan pada tahap pemodelan. Proses ini meliputi pembersihan data, transformasi, dan seleksi atribut, guna memastikan data siap untuk diproses lebih lanjut dalam tahap pemodelan.

a) Penanganan Data yang Hilang

Penanganan data yang hilang adalah aspek penting dalam analisis data yang tidak boleh diabaikan. Data yang hilang dapat muncul karena berbagai alasan, termasuk kesalahan pengukuran, data yang tidak lengkap, atau kehilangan data selama proses pengumpulan. Pembahasan ini mencakup berbagai teknik untuk menangani data yang hilang, mulai dari metode sederhana seperti penghapusan data dan imputasi mean, hingga pendekatan yang lebih kompleks seperti penggunaan model statistik dan algoritma pembelajaran mesin. Tujuan utama adalah untuk memahami kelebihan dan kekurangan masing-masing metode, serta dampaknya terhadap hasil analisis dan kesimpulan yang ditarik dari penelitian ini. Dengan penanganan yang

tepat, kita dapat meminimalkan bias dan memastikan keakuratan serta keandalan hasil analisis.

b) Penanganan *Outlier*

Penanganan *outlier* adalah langkah penting dalam analisis data, karena *outlier* dapat secara signifikan mempengaruhi hasil analisis dan mengarah pada kesimpulan yang menyesatkan. *Outlier* adalah titik data yang berbeda jauh dari mayoritas data lainnya, dan dapat terjadi karena kesalahan pengukuran, anomali dalam data, atau variasi yang sebenarnya. Dalam pembahasan ini, kita akan mengevaluasi berbagai metode untuk mendeteksi dan menangani *outlier*, termasuk analisis panjang teks, analisis frekuensi kata, serta penggunaan statistik *z-score*. Menggunakan metode *z-score* (deviasi standar) untuk mendeteksi *outlier*, yaitu dengan mencari data yang berada di luar dua kali simpangan baku dari rata-rata. Dengan mengidentifikasi dan menangani outlier secara tepat, kita dapat memastikan bahwa analisis data lebih akurat dan hasilnya lebih andal. Pendekatan yang digunakan harus disesuaikan dengan karakteristik dan tujuan dataset untuk mencapai keseimbangan antara menghilangkan data yang tidak representatif dan mempertahankan integritas data yang valid.

- Menghitung Rata-rata dan Standar Deviasi
    - Rata-rata (mean):

$$\bar{x} = \frac{\sum x_i}{n}$$

Rata-rata mewakili nilai tengah dari distribusi rasio kata unik.

    - Standar deviasi (deviation):

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}}$$

Standar deviasi menunjukkan seberapa tersebar nilai rasio kata unik dari rata-rata.

- Menentukan Batas *Outliers*

Outliers didefinisikan sebagai nilai-nilai yang berada di luar dua kali standar deviasi dari rata-rata. Dua batas dihitung:

- Batas atas (*upper bound*)

$$upper\ bound = \bar{x} + 2 * s$$

Rasio yang lebih besar dari nilai ini dianggap *outlier*.

- Batas bawah (*lower bound*)

$$lower\ bound = \bar{x} - 2 * s$$

Rasio yang lebih kecil dari nilai ini juga dianggap *outlier*.

Keterangan:

$\bar{x}$: nilai rata-rata

x_i : nilai data ke-i

n : banyaknya data

s : nilai standar deviasi

- Memfilter *Outliers*

Setelah batas dihitung, setiap baris diukur:

- Jika nilai unique_word_ratio melebihi *upper bound,* baris tersebut dianggap sebagai outlier tinggi (*high outlier*).
- Jika nilai unique_word_ratio berada di bawah *lower bound*, baris tersebut dianggap sebagai outlier rendah (*low outlier*).

c) Konversi Tipe Data

Langkah penting dalam pemrosesan dan analisis data adalah konversi tipe data, terutama ketika bekerja dengan berbagai sumber data yang memiliki format yang berbeda. Proses ini melibatkan mengubah nilai tipe data, seperti mengubah *string* menjadi *integer* atau *float*, yang memungkinkan kita melakukan operasi aritmetika dan analisis statistik lebih lanjut. Konversi tipe data adalah penting dalam pengolahan data karena memastikan bahwa data disimpan dalam format yang sesuai dengan kebutuhan analisis, yang meningkatkan akurasi dan efisiensi pengolahan data. Dalam pembersihan data, teknik ini bermanfaat karena dapat mengidentifikasi dan mengkoreksi data yang tidak sesuai tipe atau format. Oleh karena itu, untuk menjamin integritas dan kesesuaian data selama proses analisis, konversi tipe data adalah langkah penting.

d) Pembersihan Data

Analisis data bertujuan untuk memastikan kualitas dan integritas data, dan proses penting yang dikenal sebagai pembersihan data melibatkan identifikasi dan penanganan data yang tidak akurat, rusak, tidak relevan, atau tidak lengkap. Dengan demikian, pembersihan data mengurangi kemungkinan kesalahan dan bias dalam analisis dan meningkatkan keandalan hasil. Berbagai metode digunakan dalam proses ini, termasuk penanganan data yang hilang, penghapusan data duplikat, koreksi kesalahan format, dan standardisasi nilai. Kita dapat memastikan bahwa data yang digunakan dalam analisis adalah representatif dan bermakna dengan menerapkan teknik pembersihan data yang tepat. Ini akan membantu kita membuat keputusan yang lebih informatif dan akurat.

Salah satu teknik yang efektif untuk membersihkan data adalah *regular expressions* (regex), terutama ketika bekerja dengan teks yang tidak terstruktur. Regex menawarkan kemampuan untuk mendeteksi dan mengubah string sesuai dengan pola tertentu, sehingga sangat berguna untuk berbagai tugas pembersihan data, seperti menghilangkan karakter yang tidak diinginkan, mengekstrak informasi khusus, dan mengubah format teks.

- Menghapus Karakter yang Tidak Relevan

Seringkali, karakter seperti tanda baca, simbol, dan *whitespace* yang berlebihan harus dihapus atau diganti untuk memastikan data konsisten. Kita dapat dengan mudah menentukan pola ini dan menghilangkannya dari dataset dengan regex.

- Validasi Format Data

Regex memungkinkan validasi format data seperti email, nomor telepon, alamat IP, dan kode pos. Dengan mendefinisikan pola regex yang tepat, kita dapat memastikan bahwa data sesuai dengan format yang diharapkan dan mengidentifikasi entri yang tidak valid.

- Ekstraksi Data

Regex dapat digunakan untuk mengekstrak informasi khusus dari teks, seperti tanggal, URL, atau angka. Ini sangat membantu ketika bekerja dengan data yang

mengandung banyak informasi tersembunyi yang harus diisolasi untuk analisis lebih lanjut.

- Penggantian dan Substitusi

Dengan menggunakan aturan tertentu, kita dapat mengganti atau menyubstitusi bagian teks dengan regex. Misalnya, ganti singkatan dengan kata lengkap atau ubah semua huruf kapital menjadi huruf kecil.

Implementasi pustaka *regex* digunakan untuk mengubah teks menjadi huruf kecil (*case folding*), proses pembersihan kata (*cleaning*) dengan menghapus angka, tanda baca dan karakter khusus, karakter non-alphanumeric, dan spasi berlebih [19]. Implementasi pembersihan data terdapat pada tabel 3.2.

Tabel 3.2 Pembersihan data dengan pustaka *regex*

Teks Asli	Merancang, mengembangkan, dan memelihara layanan *backend* dan API yang dapat diskalakan menggunakan bahasa pemrograman dan kerangka kerja yang relevan (misalnya, *Python / Django, Java / Spring*).
Case Folding	merancang, mengembangkan, dan memelihara layanan *backend* dan *api* yang dapat diskalakan menggunakan bahasa pemrograman dan kerangka kerja yang relevan (misalnya, p*ython / django, java / spring*).
Cleaning	merancang mengembangkan dan memelihara layanan backend dan api yang dapat diskalakan menggunakan bahasa pemrograman dan kerangka kerja yang relevan misalnya python django java spring

e) *Stopword Removal*

Stopword removal bertujuan menyesuaikan teks mentah menjadi format yang sesuai untuk penelitian, akan berfokus pada pembagian dan penghapusan elemen yang tidak dibutuhkan. Tahap *stopword removal* menggunakan bahasa pemrograman Python dengan pustaka spaCy. Proses yang dilakukan *stopword removal* (penghapusan kata), *tokenizing* (pemecahan kata), *stemming* (mengurai kata dasar), dan *join* (menggabungkan kembali) [19]. Implementasi transformasi data terdapat pada tabel 3.3.

Tabel 3.3 Transformasi data dengan pustaka *spacy*

Stopword Removal	layanan backend api bahasa pemrograman kerangka kerja relevan python django java spring
Tokenizing	layanan || backend || api || bahasa || pemrograman || kerangka || kerja || relevan || python || django || java || spring
Join	layanan backend api bahasa pemrograman kerangka kerja relevan python django java spring

f) Seleksi Fitur

Penting untuk menentukan subset data yang relevan dan representatif untuk tahap pemodelan, langkah tersebut adalah pemilihan data. Data yang dimaksud adalah variabel *subcategory* sebagai label dan *descriptions* sebagai fitur. Pemilihan ini didasarkan pada kriteria kecenderungan deskripsi dengan kategori yang dimiliki, untuk memastikan bahwa data yang digunakan mencerminkan karakteristik yang diperlukan untuk pelatihan model.

3.3.5 *Modeling* (Tahap Pemodelan)

Tahap pemodelan ini berkaitan langsung dengan teknik *data mining*, serta penentuan algoritma yang akan digunakan. Pada tahap pemodelan merupakan proses implementasi model klasifikasi data untuk mengukur seberapa akurat model dalam memprediksi data. Pemodelan menggunkan algoritma *machine learning*, seperti KNN, NBC, dan SVM dari pustaka scikit-Learn. Pemilihan tiga algoritma, bertujuan untuk membandingkan tingkat akurasi dari masing-masing model data pada penelitian ini. Implementasi model menggunakan bahasa Python di perangkat lunak VS Code.

Klasfikasi data melibatkan pembagian set data menjadi dua, yaitu set data pelatihan *(data train)* dan set data pengujian *(data test)*. Fungsi set data pelatihan untuk memprediksi kelas model baru yang belum pernah dilihat sebelumnya, sedangkan set data pengujian untuk mengevaluasi tingkat akurasi model dalam memprediksi data. Pembagian set data dilakukan secara acak dengan proporsi yang telah ditentukan, 80% untuk set data pelatihan dan 20% untuk set data pengujian.

3.3.6 *Evaluation* (Tahap Evaluasi)

Pada tahapan ini melakukan evaluasi model secara menyeluruh dan meninjau langkah-langkah yang dijalankan untuk membangun model serta memastikannya mencapai tujuan penelitian. Tahap evaluasi merupakan langkah tambahan yang dilakukan untuk mengevaluasi secara mendalam tujuan *data mining* dalam melakukan pemodelan sesuai dengan yang diinginkan. Tahap evaluasi sangat penting untuk mengukur kinerja model yang telah dikembangkan.

Dalam kasus ini, algoritma KNN, NBC dan SVM dievaluasi berdasarkan nilai *evaluation measure*, *evaluation measure* dengan parameter pengukuran *accuracy, precision, recall,* dan *f1-score.* Validasi digunakan untuk menentukan jenis model terbaik melalui *confusion matrix* [29] sebagai informasi mengenai hasil klasifikasi aktual yang dapat diprediksi oleh sistem melalui nilai *accuracy, precision, recall,* dan *f1-score*. Evaluasi dilakukan dengan membandingkan model set pelatihan dengan set pengujian.

Accuracy, atau akurasi, adalah metrik yang digunakan untuk mengevaluasi seberapa baik model klasifikasi dalam memprediksi label yang tepat. Akurasi

dihitung sebagai rasio antara jumlah prediksi yang benar dengan jumlah total prediksi yang dibuat. Metrik ini memberikan gambaran umum tentang seberapa efektif model dalam membuat prediksi yang benar.

$$accuracy = \frac{TP + TN}{P + N}$$

Precision, atau presisi dan *recall*, atau penarikan juga banyak digunakan dalam klasifikasi. Presisi mengukur berapa banyak dari prediksi positif yang benar-benar positif. Presisi dapat dianggap sebagai ukuran ketepatan, yaitu persentase tupel yang diberi label positif yang sebenarnya memang termasuk dalam kelas positif. Presisi mengukur sejauh mana prediksi positif model benar-benar akurat, dihitung dengan rumus,

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

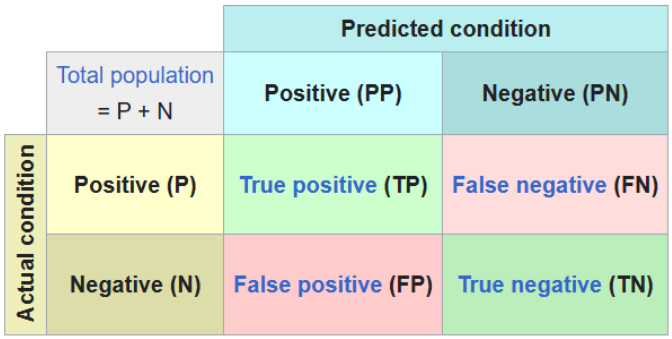
sedangkan *recall* adalah ukuran kelengkapan, yaitu persentase tupel positif yang benar-benar diberi label positif oleh model. *Recall* mengukur seberapa banyak dari kasus positif yang sebenarnya dikenali oleh model, dihitung dengan rumus.

$$recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

F1-score, F adalah rata-rata harmonik presisi dan *recall*. Ini memberikan bobot yang sama untuk presisi dan penarikan. Ini memberikan gambaran keseimbangan antara keduanya [29].

$$F1 - score = \frac{2\ precision\ x\ recall}{precision + recall}$$

Confusion Matrix untuk mengukur seberapa akurat model SVM dalam memodelkan klasfikasi. Hasil evaluasi digunakan untuk mengukur seberapa baik model dapat memberikan prediksi yang akurat dan relevan dengan tujuan bisnis, serta untuk mengidentifikasi area-area yang perlu diperbaiki atau dioptimalkan dalam model tersebut.



Gambar 3.1 Formula *confusion matrix.*

Catatan: TP, TN, FP, P, dan N mengacu pada jumlah *true positive, true negative, false positive, positive, and negative samples, respectively.*

Proses evaluasi dilakukan dengan menghitung matriks evaluasi, seperti nilai-nilai *accuracy, recall,* dan *F1-score* dihitung dengan menggunakan hasil *confusion matrix*. Ini menunjukkan seberapa baik model mengklasifikasikan data. Kemudian analisis *confusion matrix,* analisis ini digunakan untuk memahami kesalahan klasifikasi, seperti kesalahan dalam memprediksi kelas positif atau negatif. Analisis ini membantu dalam menentukan area di model yang perlu diperbaiki.

3.3.7 *Deployment* (Tahap Pemodelan)

Model klasifikasi teks yang telah dibangun dan dievaluasi diterapkan pada tahap penerapan. Lingkungan ini memungkinkan pengguna berinteraksi dengan hasil analisis melalui antarmuka yang mudah digunakan. Alat yang tepat untuk tujuan ini adalah Google Looker Studio. Looker Studio memungkinkan membuat visualisasi data interaktif dan dashboard yang menampilkan hasil model klasifikasi dengan cara yang informatif dan mudah dipahami [37].

a. Integrasi Data ke Looker Studio: Proses *deployment* dimulai dengan mengimpor data hasil prediksi dari model ke Looker Studio. *Dataframe* prediksi diimpor ke Looker Studio dengan menghubungkannya ke sumber data seperti file CSV Google Sheets. Ini memastikan bahwa data yang ditampilkan dalam visualisasi adalah yang paling relevan dan mutakhir.
b. Pembuatan *Dashboard*: Setelah data diintegrasikan dengan baik, langkah selanjutnya adalah membuat *dashboard* yang menampilkan hasil model secara visual. Beberapa komponen utama termasuk dalam *dashboard* ini:

Diagram: Untuk memberikan gambaran umum tentang proporsi setiap kelas dalam hasil model, diagram pie menunjukkan distribusi label prediksi. Diagram batang menunjukkan perbandingan jumlah prediksi untuk masing-masing label, membantu dalam menganalisis frekuensi dan distribusi hasil klasifikasi.

Visualisasi Teks: Untuk memberikan informasi lebih lanjut tentang konten data yang diklasifikasikan, teks yang telah dikelompokkan berdasarkan label prediksi ditampilkan.

c. Interaktivitas dan Analisis: Looker Studio memungkinkan penggunaan elemen *dashboard* interaktif seperti filter dan kontrol waktu. Ini memungkinkan pengguna menyesuaikan tampilan data sesuai dengan kebutuhan analisis mereka. Fitur ini memungkinkan pengguna melihat hasil klasifikasi dalam konteks yang lebih spesifik dan mendapatkan pemahaman yang lebih baik tentang data.
d. Publikasi dan Akses: Publikasi adalah langkah terakhir setelah pembuatan dan verifikasi *dashboard* selesai. Dengan Looker Studio, pengguna dapat berbagi *dashboard* melalui link langsung atau dengan mengundang pengguna tertentu. *Dashboard* tersedapat diintegrasikan ke dalam aplikasi internal atau situs *web* sehingga pengguna yang membutuhkan informasi hasil klasifikasi dapat mengaksesnya [28].

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan, studi ini mencapai beberapa kesimpulan berikut:

1. Data yang memiliki keseragaman tipe data, dan format saat pengolahan data dapat meningkatkan skor akurasi pada proses pemodelan. Data yang digunakan dalam penelitian terdiri dari 2340 baris lowongan pekerjaan, direpresentasikan dengan baik melalui pengolahan data seperti pembersihan dengan EDA, vektorisasi, dan seleksi fitur. Hal ini memastikan model yang dihasilkan relevan dan dapat diandalkan.
2. Pendekatan CRISP-DM memungkinkan proses penelitian yang terstruktur, dimulai dari pemahaman masalah, persiapan data, hingga *deployment* hasil akhir. *Framework* ini efektif dalam mendukung analisis dan pemodelan data mining berbasis *machine learning*.
3. Berdasarkan pemodelan yang telah dilakukan model SVM memperoleh tingkat akurasi terbaik dengan skor 86.75%. Skor ini menunjukkan bahwa model memiliki performa kinerja sangat baik saat memprediksi label. Kemudian diikuti oleh KNN dengan skor 83.33%, dan NBC dengan skor terendah 79.49%.
4. Visualisasi data menggunakan Looker Studio memberikan kemudahan bagi pengguna dalam memahami hasil analisis. Visualisasi seperti menampilkan *sampling dataset*, visualisasi label serta visualisasi label dengan persentase.
5. Hasil prediksi model memperoleh label berupa sub-kategori bidang teknologi dan informasi, dengan total 21 *value* beserta skill pendukung untuk masing-masing lowongan. Peringkat pertama dengan total 34.1%

berlabel *Business/System Analyst*, kedua 22.6% berlabel *Network & System Administration*, dan ketiga 8% berlabel *Developer/Programmer*. Oleh karena itu, dapat disimpulkan lowongan pekerjaan dengan kebutuhan terbanyak adalah *Business/System Analyst.*

6. Penelitian ini berhasil mengidentifikasi berbagai keterampilan yang relevan untuk masing-masing subkategori pekerjaan di bidang teknologi informasi, seperti "*management*", "*server*", "*security*", "*application*", dan "*design*". Temuan ini dapat menjadi acuan bagi institusi pendidikan dan pelamar kerja dalam menyesuaikan keterampilan mereka.
7. Dengan menerapkan *data mining*, penelitian ini memberikan wawasan berharga tentang kebutuhan pasar kerja, mengurangi waktu yang dibutuhkan untuk mencocokkan keterampilan pelamar dengan deskripsi pekerjaan, serta meningkatkan efisiensi rekrutmen di industri teknologi informasi. Penelitian ini juga menyoroti urgensi universitas untuk menyesuaikan kurikulum dengan kebutuhan pasar kerja, khususnya dalam konteks bidang teknologi informasi. *Machine learning* terbukti efektif dalam mengidentifikasi tren keterampilan yang berkembang dan membantu mahasiswa menghadapi tantangan di era digital. Harapannya penelitian ini dapat memberikan kontribusi pada pemahaman tren keterampilan yang dibutuhkan oleh perusahaan serta membantu pencari kerja dalam merencanakan karir mereka.

5.2 Saran

Sebagai ide pengembangan ke depan untuk penelitian akademis, seluruh ide tertulis dalam saran. Saran tersebut terlampir sebagai berikut.

1. Penelitian mendatang disarankan untuk mengambil data secara berkala selama satu tahun dengan cakupan sektor industri yang lebih luas, selain teknologi informasi. Hasilnya dapat divisualisasikan berdasarkan tren bulanan untuk mengevaluasi generalisasi model klasifikasi sekaligus memahami pola permintaan kerja di berbagai sektor industri.

2. Penelitian mendatang dapat memanfaatkan data lowongan pekerjaan *real-time* untuk mengidentifikasi tren keterampilan yang terus berkembang di pasar kerja, sehingga hasilnya lebih relevan dan *up-to-date*.
3. Disarankan untuk mengintegrasikan teknik pemrosesan teks yang lebih canggih, seperti *word embeddings* (contohnya Word2Vec atau BERT), yang dapat meningkatkan representasi semantik dari data teks.
4. Gunakan operator SMOTE, supaya performa model memiliki kinerja klasifikasi lebih baik. Serta disarankan untuk mengeksplorasi algoritma *machine learning* atau *deep learning* lainnya, seperti Random Forest, Gradient Boosting, atau Neural Networks, untuk membandingkan performa mereka dengan algoritma yang telah digunakan dalam penelitian ini.
5. Penelitian selanjutnya dapat mengaplikasikan model yang telah dibangun ke dalam sistem rekrutmen nyata untuk mengevaluasi efektivitasnya dalam membantu perusahaan mencocokkan pelamar kerja dengan lowongan pekerjaan.
6. Selain Looker Studio, disarankan untuk mengeksplorasi alat visualisasi data lainnya, seperti Tableau atau Power BI, untuk memberikan fleksibilitas dan kemampuan visualisasi yang lebih kaya.

DAFTAR PUSTAKA

[1] Wikipedia, "JobStreet," Wikipedia. Accessed: Feb. 05, 2024. [Online]. Available: https://en.wikipedia.org/wiki/JobStreet#cite_note-1

[2] E. Seek, "5 Alasan JobStreet Merupakan Pilihan Utama Perekrut Kerja," Employer Seek. Accessed: Feb. 05, 2024. [Online]. Available: https://id.employer.seek.com/id/market-insights/article/5-alasan-jobstreet-merupakan-pilihan-utama-perekrut-kerja

[3] A. Zhang, *Data Analytics: Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life*. North Charleston: CreateSpace Independent Publishing Platform, 2017. doi: https://dl.acm.org/doi/book/10.5555/3153180.

[4] V. K. Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, *Introduction to Data Mining*. 2005.

[5] M. Hofmann and R. Klinkenberg, *Data Mining and Knowledge Discovery Series*. 2014.

[6] A. M. Zuhdi, E. Utami, and S. Raharjo, "ANALISIS SENTIMENT TWITTER TERHADAP CAPRES INDONESIA 2019 DENGAN METODE K-NN," *J. Inf. Politek. Indonusa Surakarta*, vol. 5, pp. 1–7, 2019.

[7] V. Nurcahyawati and Z. Mustaffa, "Improving sentiment reviews classification performance using support vector machine-fuzzy matching algorithm," *Bull. Electr. Eng. Informatics*, vol. 12, no. 3, pp. 1817–1824, 2023, doi: 10.11591/eei.v12i3.4830.

[8] N. S. Wardani, A. Prahutama, and P. Kartikasari, “Analisis Sentimen Pemindahan Ibu Kota Negara Dengan Klasifikasi Naïve Bayes Untuk Model Bernoulli Dan Multinomial,” *J. Gaussian*, vol. 9, no. 3, pp. 237–246, 2020, doi: 10.14710/j.gauss.v9i3.27963.

[9] D. Janeth, L. Cuesta, R. Jairo, and R. Jairo, “Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test,” *Procedia Comput. Sci.*, vol. 198, no. 2020, pp. 512–517, 2022, doi: 10.1016/j.procs.2021.12.278.

[10] C. Pete *et al.*, “CRISP-DM 1.0,” *Cris. Consort.*, p. 76, 2000.

[11] Forbes, “Turning Classifieds Into Cash,” Forbes. Accessed: Feb. 05, 2024. [Online]. Available: https://www.forbes.com/global/2008/0128/014.html

[12] Budi Santosa, *Data mining : Teknik pemanfaatan data untuk keperluan bisnis / Budi Santosa*, 978th ed. Yogyakarta: Graha Ilmu, 2007.

[13] J. P. Jiawei Han, Micheline Kamber, *Data Mining - Concepts And Techniques, 3rd Edition*. Amsterdam: Kaufmann-Elsevier, 2012.

[14] A. Géron, *Hands-On Machine Learning with Scikit-Learn*, 1st editio. Sebastopol: O’Reilly Media, Inc., 2019. doi: dl.acm.org/doi/10.5555/3378999.

[15] Y. Nurdiansyah, A. Andrianto, and L. Kamshal, “New book classification based on Dewey Decimal Classification (DDC) law using tf-idf and cosine similarity method,” *J. Phys. Conf. Ser.*, vol. 1211, no. 1, 2019, doi: 10.1088/1742-6596/1211/1/012044.

[16] J. P. Jiawei Han, Micheline Kamber, *Data mining : concepts and techniques*, 3rd editio. Waltham,: Morgan Kaufmann/Elsevier, Waltham, MA, ©2012, 2012. doi: https://dl.acm.org/doi/book/10.5555/1972541.

[17] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, “DMME : Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model,” vol. 79, no. March, pp. 403–408, 2023.

[18] R. Ribeiro, A. Pilastri, C. Moura, F. Rodrigues, R. Rocha, and P. Cortez, "Predicting the tear strength of woven fabrics via automated machine learning: An application of the CRISP-DM methodology," *ICEIS 2020 - Proc. 22nd Int. Conf. Enterp. Inf. Syst.*, vol. 1, pp. 548–555, 2020, doi: 10.5220/0009411205480555.

[19] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[20] Y. A. Singgalen, "Analisis Sentimen Wisatawan terhadap Kualitas Layanan Hotel dan Resort di Lombok Menggunakan SERVQUAL dan CRISP-DM," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 1870–1882, 2023, doi: 10.47065/bits.v4i4.3199.

[21] A. Geron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, Second edi. Sebastopol: O'Reilly Media, Inc., 2019.

[22] Z. Luo *et al.*, "VSCode: General Visual Salient and Camouflaged Object Detection with 2D Prompt Learning," pp. 17169–17180, 2023, doi: 10.1109/CVPR52733.2024.01625.

[23] B. Johnson, *Visual Studio Code: End-to-End Editing and Debugging Tools for Web Developers*. John Wiley & Sons, Inc, 2019. doi: 10.1002/9781119588238.

[24] A. Kowalczyk, *Support Vector Machines Succinctly*. Morrisville: Syncfusion, 2017. [Online]. Available: https://www.dbooks.org/support-vector-machines-succinctly-5591635185/read/

[25] A. Downey, J. Elkner, and C. Meyers, "Think Python: How to Think Like a Computer Scientist Πώς να Σκέφτεσαι σαν Επιστήμονας της Πληροφορικής," p. 304, 2014.

[26] A. Sweigart, *Automate the Boring Stuff with Python, 2nd Edition: Practical Programming for Total Beginners*. No Starch Press, 2019. [Online].

Available: https://automatetheboringstuff.com/#toc

[27] S. Interactive, “Google Data Studio: What It Is & How to Use It,” Seer Interactive. [Online]. Available: https://www.seerinteractive.com/insights/google-data-studio-whats-working-whats-missing

[28] G. Inc., “Connect and visualize all your data in Looker Studio.” Accessed: Feb. 29, 2024. [Online]. Available: https://codelabs.developers.google.com/codelabs/community-connectors/#0

[29] L. Hurst, *Hands On With Google Data Studio: a Data Citizens Survival Guide*. Canada: John Wiley & Sons, Inc, 2020. doi: doi.org/10.1002/9781119616238.

[30] R. Wirth, “CRISP-DM : Towards a Standard Process Model for Data Mining,” no. 24959.

[31] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, “DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model,” *Procedia CIRP*, vol. 79, no. March, pp. 403–408, 2019, doi: 10.1016/j.procir.2019.02.106.

[32] M. S. Kanish Shah, Henil Patel, Devanshi Sanghvi, “A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification,” *Augment. Hum. Res.*, 2020, doi: doi.org/10.1007/s41133-020-00032-0.

[33] A. Hermawan, I. Jowensen, J. Junaedi, and Edy, “Implementasi Text-Mining untuk Analisis Sentimen pada Twitter dengan Algoritma Support Vector Machine,” *JST (Jurnal Sains dan Teknol.*, vol. 12, no. 1, pp. 129–137, 2023, doi: 10.23887/jstundiksha.v12i1.52358.

[34] A. Khumaidi, “Data Mining for Predicting the Amount of Coffee Production Using Crisp-Dm Method,” *J. Techno Nusa Mandiri*, vol. 17, no. 1, pp. 1–8, 2020, doi: 10.33480/techno.v17i1.1240.

[35] A. D. Adhi Putra, “Sentiment Analysis on User Reviews of the Bibit and Bareksa Application with the KNN Algorithm,” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 636–646, 2021.

[36] E. Kristoffersen, O. O. Aremu, F. Blomsma, P. Mikalef, and J. Li, *Exploring the Relationship Between Data Science and Circular Economy: An Enhanced CRISP-DM Process Model*, vol. 11701 LNCS. Springer International Publishing, 2019. doi: 10.1007/978-3-030-29374-1_15.

[37] S. Bonelli, “What is Google’s Looker Studio and how you can use it,” Search Engine Land. Accessed: Feb. 12, 2024. [Online]. Available: https://searchengineland.com/google-looker-studio-258871

[38] Hendro, “Ari Kuncoro: Hambatan Universitas dalam Merumuskan Kurikulum Baru,” *21 Juni*, 2020. [Online]. Available: https://feb.ui.ac.id/2019/06/21/ari-kuncoro-hambatan-universitas-dalam-merumuskan-kurikulum-baru/