IMPLEMENTASI ALGORITMA *SUPPORT VECTOR MACHINE* PADA SISTEM KLASIFIKASI *SUBJECT* SKRIPSI MAHASISWA UNIVERSITAS LAMPUNG

(Skripsi)

Oleh
RIZKY HADI
NPM 1815061002

FAKULTAS TEKNIK

UNIVERSITAS LAMPUNG

BANDAR LAMPUNG

2023

ABSTRAK

IMPLEMENTASI ALGORITMA *SUPPORT VECTOR MACHINE* PADA SISTEM KLASIFIKASI *SUBJECT* SKRIPSI MAHASISWA UNIVERSITAS LAMPUNG

Oleh
RIZKY HADI

Unit Pelaksana Teknis (UPT) Perpustakaan Universitas Lampung bertanggung jawab terhadap pengarsipan karya ilmiah mahasiswa (skripsi). Pengarsipan dilakukan secara manual, dengan menggunakan sistem klasifikasi DDC atau *Dewey Decimal Classification,* sehingga memungkinkan adanya kecenderungan ketidaktepatan pemilihan *Subject* dan durasi pengelompokkan yang lama. Penelitian ini bertujuan untuk menerapkan algoritma *Support Vector Machine* (SVM) dalam mengklasifikasi *subject* skripsi yang tersimpan di *repository* karya akhir mahasiswa Unila. Penerapan algoritma SVM ini menggunakan metode *machine learning life cycle* yang meliputi proses *data collection, pre-processing data, data splitting, model training*, hingga proses *model evaluation.* Data yang digunakan adalah judul skripsi mahasiswa Unila yang berjumlah 1707 data kategori utama yang tersedia pada *repository*. Hasil dari penelitian ini adalah sebuah program klasifikasi *subject* skripsi berdasarkan judul dengan model SVM dimana akurasi *data training* pada proses *model training* sebesar 0,95, dan *data testing* pada proses *model evaluation* dengan tingkat akurasi 0,65, presisi 0,54, dan *recall* 0,45.

Kata kunci: Klasifikasi, DDC, *Artificial Intelligence, Machine Learning, Supervised Learning, Support Vector Machine.*

ABSTRACT

IMPLEMENTATION OF SUPPORT VECTOR MACHINE ALGORITHM IN SUBJECT CLASSIFICATION SYSTEM THESIS STUDENTS OF LAMPUNG UNIVERSITY

BY
RIZKY HADI

The Technical Implementation Unit (UPT) of the University of Lampung Library is responsible for archiving student scientific work (thesis). Filing is done manually, using the DDC or Dewey Decimal Classification system, thus allowing for the tendency of inaccurate Subject selection and long grouping durations. This study aims to apply the Support Vector Machine (SVM) algorithm in classifying thesis subjects stored in the Unila student final work repository. The application of the SVM algorithm uses the machine learning life cycle method which includes the data collection process, data pre-processing, data splitting, model training, to the model evaluation process. The data used are Unila student thesis titles totaling 1707 data. The results of this study are a thesis subject classification program based on the title with the SVM model where the accuracy of the training data in the model training process is 0.95, and data testing in the evaluation mode process with an accuracy rate of 0.65, precision 0.54, and recall 0,45.

Keywords: *Classification*, DDC, *Artificial Intelligence, Machine Learning, Supervised Learning, Support Vector Machine*

IMPLEMENTASI ALGORITMA *SUPPORT VECTOR MACHINE* PADA SISTEM KLASIFIKASI *SUBJECT* SKRIPSI MAHASISWA UNIVERSITAS LAMPUNG

Oleh

Rizky Hadi

Skripsi

Sebagai Salah Satu Syarat untuk Mendapat Gelar

SARJANA TEKNIK

Pada

Program Studi Teknik Informatika

Jursan Teknik Elektro

Fakultas Teknik Universitas Lampung

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

Judul	:	**IMPLEMENTASI ALGORITMA *SUPPORT VECTOR MACHINE* PADA SISTEM KLASIFIKASI *SUBJECT* SKRIPSI MAHASISWA UNIVERSITAS LAMPUNG**
Nama Mahasiswa	:	**Rizky Hadi**
Nomor Pokok Mahasiswa	:	**1815061002**
Program Studi	:	**Teknik Informatika**
Fakultas	:	**Teknik**

MENYETUJUI

1. Komisi Pembimbing

Meizano Ardhi Muhammad, S.T.,M.T.
NIP. 198105282012121001

Puput Budi Wintoro, S. Kom, M.T.I.
NIP. 198410312019031004

2. Mengetahui

Ketua Jurusan
Teknik Elektro

Herlinawati, S.T.,M.T.
NIP. 197103141999032001

Ketua Program Studi
Teknik Informatika

Mona Arif Muda, S.T., M.T.
NIP. 197111122000031002

MENGESAHKAN

1. Tim Penguji

Ketua : **Meizano Ardhi Muhammad, S.T.,M.T.**

Sekretaris : **Puput Budi Wintoro, S. Kom, M.T.I.**

Penguji : **Ing. Hery Dian Septama, S.T.**

2. Dekan Fakultas Teknik

Dr. Eng. Helmy Fitriawan, S.T., M.Sc.

NIP. 197509282001121002

Tanggal Lulus Ujian Skripsi: **20 Januari 2023**

SURAT PERNYATAAN

Saya yang bertandatangan dibawah ini, menyatakan bahwa skripsi saya yang berjudul "IMPLEMENTASI ALGORITMA *SUPPORT VECTOR MACHINE* PADA SISTEM KLASIFIKASI *SUBJECT* SKRIPSI MAHASISWA UNIVERSITAS LAMPUNG" dengan ini menyatakan bahwa skripsi saya dibuat oleh saya sendiri. Semua hasil 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, 4 Februari 2023

Pembuat Pernyataan,

Rizky Hadi
NPM 1815061002

RIWAYAT HIDUP

Penulis bernama Rizky Hadi yang merupakan anak tunggal dari pasangan Ahmad Beni dan Aisyah Ria Sari. Penulis lahir di Bandar Lampung pada tanggal 25 Februari 2000. Penulis telah menyelesaikan sekolah dasar di SD Negeri 1 Langkapura pada tahun 2012. SMP Islam Terpadu Fitrah Insasni pada tahun 2015 dan SMA Negeri 3 Bandar Lampung pada tahun 2018. Penulis melanjutkan pendidikan di Fakultas Teknik Jurusan Teknik Elektro Program Studi Teknik Informatika Universitas Lampung pada tahun 2018. Selama menempuh pendidikan di Jurusan Teknik Elektro Program Studi Teknik Informatika Universitas Lampung, penulis juga aktif mengikuti beberapa kegiatan sebagai berikut

1. Peserta Studi Independent Kampus Merdeka pada program Data and Artificial Intelligence tahun 2021.
2. Peserta Pelatihan FGA dijalur karir Artificial Intelligence for Junior Developer (Huawei) pada tahun 2022.
3. Anggota UKM-U Sains dan Teknologi Universitas Lampung pada tahun 2021
4. Pada bulan Februari - Maret 2021, penulis mengikuti Kuliah Kerja Nyata selama 40 hari di Langkapura Baru, Kecamatan Kemiling, Kota Bandar Lampung.

PERSEMBAHAN

Bismillahirrahmanirrahim….

Segala Puji syukur kepada Allah SWT berkat karunia, kesehatan, rahmat serta hidayah-Nya yang telah diberikan, shalawat teriring salam kepada Nabi Muhammad SAW, suri tauladan Akhlaqul Kharimah yang kita nantikan syafa'atnya di hari akhir kelak. Dengan segala kerendahan hati, saya persembahkan skripsi ini kepada:

Ibuku, yang telah melahirkanku, merawatku, membesarkanku, dan yang telah sepenuh hati mendidikku.

Ayahku tercinta, yang telah membesarkanku dengan seluruh kasih dan sayangnya, memberikan pengetahuannya, dan selalu mendukung serta mendoakan untuk keberhasilanku.

Serta, almamater yang saya sangat banggakan

UNIVERSITAS LAMPUNG

SANWACANA

Segala puji hanya bagi Allah SWT, Tuhan semesta alam yang Maha Pengasih lagi Maha Penyayang, atas limpahan rahmat, taufik serta hidayah-Nya sehingga penulis dapat menyelesaikan penulisan skripsi yang berjudul: IMPLEMENTASI ALGORITMA *SUPPORT VECTOR MACHINE* PADA SISTEM KLASIFIKASI *SUBJECT* SKRIPSI MAHASISWA UNIVERSITAS LAMPUNG. Sebagai salah satu syarat untuk memperoleh gelar Sarjana Teknik pada Fakultas Teknik Universitas Lampung. Dalam penyelesaian skripsi ini, penulis mendapatkan banyak bantuan, bimbingan, saran dan dukungan dari segenap pihak, baik secara langsung maupun tidak langsung sehingga penyusunan skripsi ini berjalan dengan baik. Maka pada kesempatan ini penulis mengucapkan terima kasih kepada:

1. Bapak Dr. Eng. Helmy Fitriawan, S.T., M.Sc selaku Dekan Fakultas Teknik Universitas Lampung.
2. Ibu Herlinawati, S.T.,M.T. selaku Ketua Jurusan Teknik Elektro Universitas Lampung.
3. Bapak Mona Arif Muda, S.T., M.T. selaku Ketua Program Studi Teknik Informatika Jurusan Teknik Elektro Universitas Lampung.
4. Bapak Meizano Ardhi Muhammad, S.T.,M.T. selaku Pembimbing Utama, yang telah bersedia meluangkan waktu untuk memberikan pengarahan dan bimbingan dalam pembuatan skripsi ini hingga selesai.
5. Bapak Puput Budi Wintoro, S. Kom, M.T.I. selaku Pembimbing Kedua, yang telah bersedia meluangkan waktu untuk memberikan pengarahan dan bimbingan dalam pembuatan skripsi ini hingga selesai.
6. Pak Wahyu Eko Sulistiono, S.T., M.Sc, selaku Pembimbing Akademik yang telah bersedia meluangkan waktu untuk memberikan pengarahan dan bimbingan selama menempuh pendidikan Program Studi Teknik Informatika Universitas Lampung
7. Bapak Rio Ariestia P, S. Kom. M.T.I. selaku Pembimbing Bayangan, yang telah bersedia memberikan pengarahan dalam pembuatan skripsi ini hingga selesai.

8. Seluruh Dosen Program Studi Teknik Informatika yang telah membagikan ilmunya kepada penulis.
9. Seluruh teman teman Teknik Informatika Angkatan 2018 selaku teman kelas.
10. Teman sekaligus mentor saya Mazi Prima Reza yang selalu sabar dalam mengajar saya
11. Sahabat saya Irvani Andreas, Emir Rasyid, Dea Elfira yang telah mendukung, menghibur, dan membantu saya selama proses pembuatan skripsi ini.

Semoga Allah SWT membalas segala bentuk kebaikan hati dan jasa yang telah kalian berikan kepada saya. Saya menyadari meskipun skripsi ini sudah disusun dengan sebaik mungkin, skripsi ini masih terdapat kekurangan dan masih jauh dari kata sempurna, namun saya sangat berharap melalui skripsi ini akan memberikan manfaat bagi siapapun yang membacanya dan bagi penulis dalam mengembangkan dan mengamalkan ilmu pengetahuan yang telah ditempuh selama ini.

Bandar Lampung, 4 Februari 2023

Penulis,

Rizky Hadi
NPM 1815061003

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1. 1 Latar Belakang

Unit Pelaksana Teknis (UPT) Perpustakaan Universitas Lampung bertanggung jawab terhadap pengarsipan hasil penelitian yang dilakukan oleh sivitas akademika perguruan tinggi Universitas Lampung. Koleksi pengarsipan yang biasa dilakukan diantaranya adalah buku, tesis, hasil penelitian dosen, dan karya ilmiah mahasiswa (skripsi). Pengarsipan yang dilakukan UPT Perpustakaan Lampung menggunakan sistem yang berbasis pada teknologi *ePrint* yang bisa diakses melalui website digilib.unila.ac.id. Dalam sistem *ePrint* diperlukan pengelompokkan berdasarkan *Subject* pada penelitian yang akan diarsipkan.

Proses pengelompokkan *Subject* ini ditentukan sendiri oleh staf perpustakaan yang dipengaruhi oleh pengalaman dan wawasan masing-masing staff secara manual, dengan menggunakan sistem klasifikasi DDC atau *Dewey Decimal Classification,* sehingga memungkinkan ada kecenderungan ketidaktepatan pemilihan *Subject* dan durasi pengelompokkan yang lama. Padahal hal ini sangat mempengaruhi saat peneliti-peneliti lain membutuhkan referensi dari repositori penelitian yang dimiliki oleh perguruan tinggi pada UPT Perpustakaan Lampung, sehingga diperlukan teknologi dalam menjawab permasalahan tersebut.

Sebuah teknologi yang biasa digunakan dalam proses klasifikasi adalah *Machine Learning* (ML). ML mempelajari teknik yang dapat memberikan komputer sebuah potensial untuk belajar melalui data yang sudah diperoleh sehingga bisa melakukan tugas tertentu. ML bisa melakukan tugas yang beragam tergantung

dari apa yang ML pelajari. ML menggunakan algoritma untuk memprediksi data guna menentukan data tersebut, termasuk dalam kategori yang sudah ditentukan sebelumnya.

Dalam penelitian Osisanwo F.Y dkk pada jurnal *Supervised Machine Learning Algorithms: Classification and Comparison* membandingkan keenam model *supervised learning* dan SVM memperoleh akurasi 73%. Penelitian Parapat, I.M., dkk pada jurnal Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak model SVM memperoleh akurasi 77%, dan pada penelitan Octaviani, P.A., dkk pada jurnal penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang model SVM memperoleh akurasi 92%. Melihat hasil akurasi dari penelitian-penelitian tersebut yang cukup tinggi, peneliti akan menggunakan *Supervised Learning* dengan algoritma *Support Vector Machine* untuk melakukan klasifikasi *Subject* dari skripsi mahasiswa Universitas Lampung pada proses pembuatan sistem dalam penelitian ini untuk membantu pustakawan UPT Perpustakaan melakukan pengarsipan.

1.2 Tujuan Penelitian

Penelitian ini memiliki tujuan untuk menerapkan algoritma *Support Vector Machine* dalam mengklasifikasi *subject* skripsi yang tersimpan di *repository* karya akhir mahasiswa Unila.

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan di atas, maka masalah yang bisa peneliti rumuskan adalah :

“Bagaimana cara mengimplementasi algoritma *Support Vector Machine* pada pengklasifikasi *Subject* untuk proses pengarsipan skrispi mahasiswa?”

1.4 Batasan Masalah

Dalam penelitian ini, pembatasan masalah dari penelitian ini adalah menggunakan,

1. *Library stopword* hanya dalam bahasa Indonesia,
2. Data yang digunakan hanya skripsi dari *repository* karya ilmiah mahasiswa Unila
3. Sistem klasifikasi yang digunakan adalah DDC atau *Dewey Decimal Classification*
4. Klasifikasi yang dilakukan hanya pada kategori utama tanpa sub kategori.

1.5 Sistematika Penelitian

Dalam penelitian skripsi ini, peneliti menggunakan penelitian yang sistematik dan terbagi menjadi lima bab, sebagai berikut:

BAB 1 PENDAHULUAN

Berisi latar belakang yang mendasari penelitian ini, rumusan masalah yang dibahas dalam penelitian, tujuan penelitian, batasan masalah, dan sistematika penelitian yang diberikan.

BAB II TINJAUAN PUSTAKA

Memuat dasar-dasar teori yang menjadi landasan dalam penelitian ini, seperti, *dewey decimal classification, artificial intelligence, machine learning, supervised*

learning, web scarpping, data pre-processing, model training, dan *support vector machine,* dan penelitian terkait

BAB III METODE PENELITIAN

Berisi waktu dan tempat penelitian, alat dan bahan yang digunakan selama penelitian, serta tahapan penelitian dengan menggunakan metode *Machine Learning Life Cycle.*

BAB IV HASIL DAN PEMBAHASAN

Memuat proses proses pengolahan data yang digunakan dalam penelitian, proses pembuatan *prediction model*, dan hasil pengujian model

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan yang didapatkan dari penelitian dan saran mengenai pengembangan lebih lanjut yang bisa dilakukan untuk penelitian ini.

BAB II
TINJAUAN PUSTAKA

2.1 *Dewey Decimal Classification*

Dewey Decimal Classification atau disingkat DDC adalah bagan sistem klasifikasi dengan sistem desimal untuk membagi bidang ilmu pengetahuan yang hingga saat ini tetap diperbarui untuk tetap mengikuti perkembangan zaman (Hamakonda, T.P., 1978). DDC dibagi ke dalam 10 kelompok bidang ilmu pengetahuan dengan menggunakan angka-angka persepuluhan;

000 – 099 Karya umum

100 – 199 Filsafat

200 – 299 Agama

300 – 399 Ilmu Sosial

400 – 499 Bahasa

500 – 599 Ilmu pengetahuan murni

600 – 699 Ilmu pengetahuan terapan/teknologi

700 – 799 Seni, olahraga, hiburan

800 – 899 Kesusasteraan

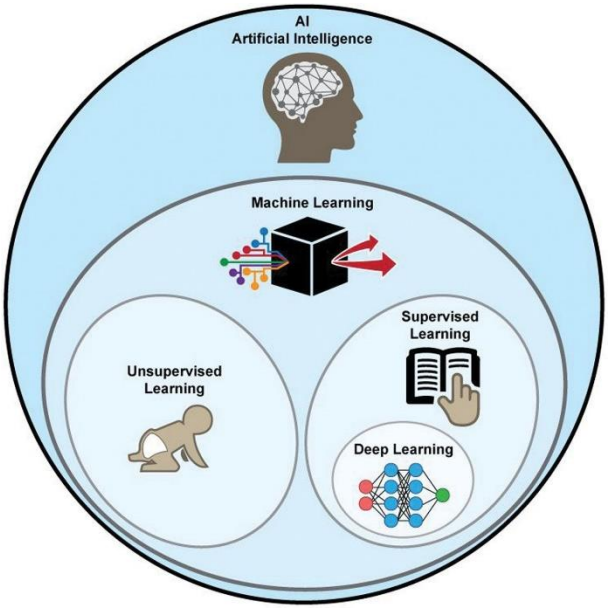
900 – 999 Biografi ilmu bumi, sejarah

2.2 *Artificial Intelligence*

Artificial Intelligence (AI) atau yang biasa disebut kecerdasan buatan adalah salah satu cabang ilmu komputer yang mempelajari bidang *smart machine* untuk memecahkan persoalan yang rumit dengan cepat untuk membantu keperluan

manusia (Widjaja. A, 2022). Beberapa unsur yang menandakan kecerdasan buatan diantaranya:

1. Adanya *system* pakar
2. Mengolah bahasa alamiah
3. Mengenali ucapan manusia dan mesin
4. Memiliki nalar dengan robotika dan sensor untuk logika
5. Mengintrepetasikan gambar dan objek melalui computer
6. Memandu atau menjadi tutor manusia
7. Memfasilitasi pembelajaran mendalam



Gambar 1. Posisi *Supervised Learning* dalam *Machine Learning*

Sumber: G. Zaharchuk, et. Al., 2018, *Deep Learning in Neuroradiology*

Pada gambar 1 menampilkan hubungan antara AI, *machine learning,* dan *deep learning* , dimana AI adalah akar dari *machine learning* dan *deep learning.* Sementara dalam *machine learning* sendiri terdapat dua teknik pembelajaran,

supervised learning dan *unsupervised learning*, yang dalam *supervised learning* terdapat pembelajaran *deep learning.*

2.3 *Machine Learning*

Machine learning (ML) merupakan sebuah ilmu computer yang mempelajari cara membuat program komputer yang belajar melalui pengalaman (Mitchell. Tom, 1998). *Machine learning* mencoba menyelesaikan masalah dengan menetapkan aturan umum pada masalah yang bervariasi, melalui penggunaan teknik statistik seperti jaringan saraf tiruan. Dalam pembuatan *machine learning* terdapat dua jenis pendekatan yaitu *supervised learning* dan *unsupervised learning.*

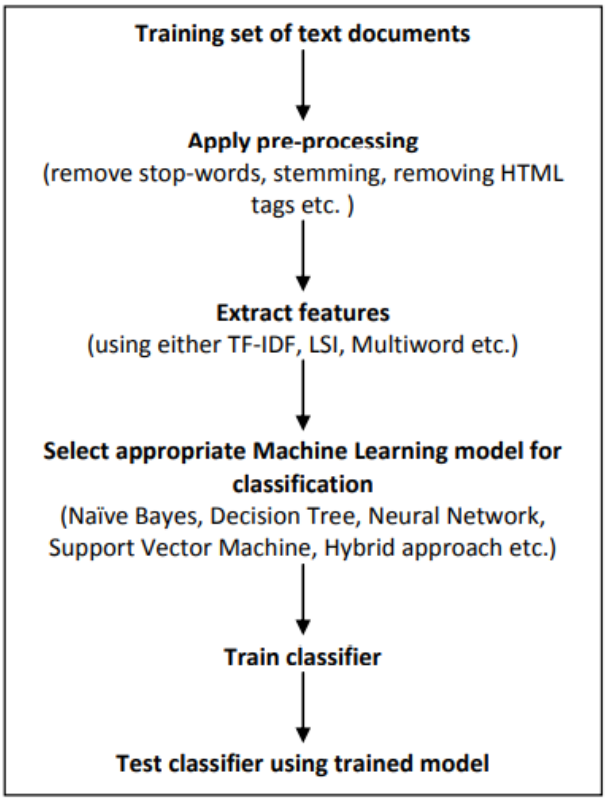
Supervised Learning adalah teknik yang umum digunakan untuk menyelesaikan masalah klasifikasi, umumnya mempelajari *output* yang diharapkan dari *input* yang belum teridentifikasi dengan menggunakan pembelajaran dari dataset yang sudah terdapat pada label di dalamnya (Vladimir Nasteski, 2017).

Supervised Learning terbagi dalam dua jenis yaitu klasifikasi dan regresi. Algoritma yang digunakan dalam proses klasifikasi memiliki beberapa algoritma populer seperti *Linear Models, Random Forest, Support Vector Machines, Naïve Bayes, Decision Tree, k-Nearest Neighbor, Neural Network, Logistic Regression,* dan *Neural Network* (Müller, A.C. and Guido, S., 2016). Disebut dengan "*supervise"* karena *machine learning* dilatih untuk mengenali dan mengidentifikasi hubungan yang mendasari koneksi antara data input dan label output. Dalam menyusun algoritma, *machine* diarahkan untuk mengolah data dan menggunakan model dengan akurasi tertinggi, untuk mendapat hasil maksimal.

2.4 Klasifikasi

Klasifikasi adalah suatu bentuk analisis data yang mengekstrak model untuk menggambarkan klasifikasi atau kategori data. Dalam klasifikasi, model dibangun untuk memprediksi label (kategorikal). Klasifikasi itu sendiri terdiri dari dua langkah atau dua proses, proses pertama adalah proses pembelajaran (proses membangun klasifikasi) dan proses kedua adalah proses klasifikasi (model dibangun untuk memprediksi label dari data yang diberikan) (Han, J,. 2011)

Pengklasifikasian teks memiliki dua cara, yaitu *clustering teks* dan klasifikasi teks (Darujati, C. 2012). *Clustering teks* betujuan untuk menemukan sebuah struktur kelompok yang belum terlihat (*unsupervised*) dari sekumpulan kalimat dalam dokumen. Sedangkan pengklasifikasian teks betujuan untuk membentuk golongan- golongan dari dokumen berdasarkan pada struktur kelompok yang sudah diketahui sebelumnya (*supervised*). Berdasarkan penjelasan ini, dapat diketahui ada beberapa cara untuk melakukan klasifikasi teks secara otomatis, yaitu pre-processing, feature extraction/selection, memilih modeling menggunakan teknik pembelajaran mesin, serta training dan testing pada classifier. (Dalal, M. K., & Zaveri, M. A., 2011).



Gambar 2. Strategi umum untuk *text classification*

Sumber: Dalal, M.K. and Zaveri, M.A., 2011. Automatic text classification: a technical review.

2.4.1 *Web Scrapping*

Web scraping adalah proses pengambilan sebuah dokumen dari internet, yang biasanya berupa halaman-halaman web dalam bahasa HTML atau XHTML untuk menganalisis dokumen tersebut yang diambil untuk digunakan kembali diperlukan protocol HTTP. HTTP atau (*Hypertext Transfer Protocol)* untuk berkomunikasi antara client dan *server*, dimana yang menjadi client adalah *web browser* atau *device* lain yang dapat mengakses konten web. Pada umumnya cara berkomunikasi antara *client* dan *server* adalah *client* melakukan *request* ke *server*, kemudian *server* mengirimkan respon terhadap *client.* Semua kegiatan tersebut diatur oleh suatu protokol yaitu protokol HTTP (Turland, M., 2010).

2.4.2 Data *pre-processing*

Pre-processing data merupakan proses mengolah data mentah atau *raw data* yang tidak berguna sebagai persiapan untuk digunakan pada proses pengolahan data (García, S., Luengo, J., Herrera, F., 2015). *Pre-processing* data dapat mengurangi ukuran dokumen teks input secara signifikan. Proses ini melibatkan aktivitas seperti *data labelling*, *data cleaning* dengan proses yang dilakukan adalah menghilangkan tanda baca, *stop-word, case folding*, dan proses akhir dari *data pre-processing* adalah *feature extraction.* Dalam *Data labelling* setiap objek yang tidak berlabel akan diberikan label berdasarkan setiap kategori data yang ada pada dataset tersebut (Cao, F., Liang, J., 2011). *Data cleaning* bekerja dengan memilih data yang relevan, data yang redundan dan *noisy* biasanya akan dihapus yang akan meningkatkan kualitas input data (García, S., Luengo, J., Herrera, F., 2015).

Stop-word adalah kata hubung yang sering muncul dalam bahasa teks. Kata hubung tidak akan bisa berdiri sendiri dalam kalimat karena tidak akan memiliki makna di dalamnya (misalnya, “dan”, “atau”, kemudian”, dan sebagainya dalam bahasa Indonesia), sehingga tidak berguna untuk klasifikasi. *Stemming* adalah tindakan mereduksi kata ke akar atau bentuk dasarnya untuk menghilangkan awalan dan akhiran dari suatu kata sehingga mengurangi kosa kata dari teks pelatihan sekitar sepertiga dari ukuran aslinya (`, M. K., & Zaveri, M. A., 2011). *Case Folding* merupakan proses untuk mengubah semua karakter alphabet pada teks menjadi huruf-huruf kecil dan menghilangkan tanda baca titik (.), koma (,), serta angka. Tujuan dari proses ini adalah mengubah setiap huruf dalam dokumen menjadi huruf kecil.

Feature extraction merujuk ke proses pemindahan operasi atribut, pemilihan atirbut, himpunan bagian dari atribut dapat digabungkan atau dapat berkontribusi pada pembuatan atribut pengganti dari data set (García, S., Luengo, J., Herrera, F., 2015). Untuk memilih atribut dari data yang berupa teks maka dilakukan metode dengan menggunakan TF-IDF *(term frequency-inverse document frequency)*. Dalam konteks klasifikasi teks, fitur atau atribut biasanya berarti kata-kata penting, banyak kata, atau frasa sering muncul dalam indikasi kategori teks. TF-IDF adalah sebuah metode yang melakukan integrasi antar *term frequency* (TF), dan *inverse document frequency* (IDF) yang berguna untuk menghitung bobot setiap kata yang digunakan dalam dokument sehingga bisa mengetahui seberapa sering suatu kata muncul dalam dokumen. TF adalah nilai yang menunjukan seberapa sering kata muncul dalam suatu dokumen. Semakin sering kata tersebut muncul semakin penting kata itu dalam dokumen. Namun, jika suatu kata terlalu sering muncul dalam *corpus* maka kata tersebut terlalu biasa dalam suatu dokumen tersebut. Hal ini diibaratkan dengan kata "atom" yang tidak biasa muncul dalam normal dokumen, tapi kata ini biasa muncul dalam dokumen fisika. Situasi ini akan diselesaikan oleh IDF. (Yoo, J.Y. and Yang, D., 2015).

Untuk menghitung IDF, rumus yang biasa digunakan adalah n/df, dimana n merupakan total dokumen yang muncul dalam *corpus*. Dalam TF-IDF yang digunakan oleh *scikit* menganggap n sebagai n+1, yang dihitung menjadi $(n+1)/df$, dan menambah 1 sebagai hasil akhir. Perhitungan TF-IDF dalam *scikit-learn* adalah sebagai berikut (Lavin, M., 2019);

$$\mathrm{TF} = \frac{f_{\mathrm{d}}(i)}{\max f_{\mathrm{d}}(j)}$$

$$\text{IDF} = \log\left(\frac{N}{df(i)+1}\right)$$

Dimana:

f_d (i) = jumlah *term* (i) pada dokumen (j)

f_d (j) = jumlah *term* dalam dokumen (j)

N = jumlah dokumen dalam *corpus*

df(i) = jumlah dokumen yang mengandung *term* (i), penambahan satu dilakukan jika tidak ditemukan df(i) dalam *corpus*

Setelah IDF ditemukan TF akan dikalikan dengan IDF

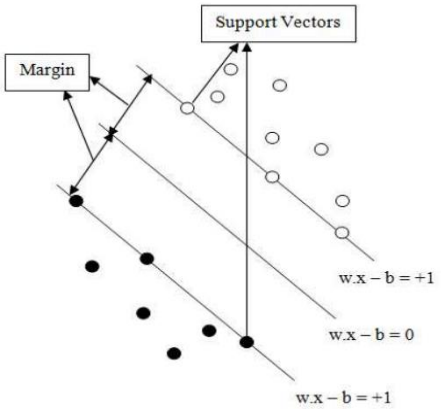
$$\text{TF-IDF} = \text{tf x idf}$$

2.4.3 *Model Training*

Model training merupakan proses *training* pada algoritma atau model yang digunakan untuk mendapatkan hasil prediksi dengan menggunakan data yang sudah melalui proses *data cleaning.* Proses ini dimulai dengan melakukan pembagian data atau *data splitting.* Tidak ada jumlah pasti dalam penentuan pembagian jumlah data untuk proses ini. Dalam tulisannya *"Recommending Training Set Sizes for Classification"*, Philip Koshute menggunakan 80% data untuk melakukan training pada 100 model data. Dengan menggunakan algoritma *Support Vector Machine* untuk mengolah data training.

SVM atau *support vector machine* adalah sistem pembelajaran yang menggunakan ruang hipotetis berupa fungsi linier dalam ruang fitur berdimensi

tinggi, yang dilatih dengan algoritma pembelajaran berdasarkan teori optimasi dengan menerapkan pembelajaran mesin yang diturunkan dari teori pembelajaran statistik. Konsep klasifikasi dengan SVM adalah mencari *hyperplane* terbaik untuk memisahkan dua kelas data dan menggunakan pendekatan support vector. (Cristianini, N. and Shawe-Taylor, 2000). *Support Vector Machine* (SVM) merupakan salah satu yang terbaik dari beberapa algoritma *machine learning* yang ada. SVM banyak digunakan untuk menyelesaikan masalah klasifikasi pola seperti *image recognition, speech recognition, text categorization, face detection, faulty card detection* dan lain sebagainya. Pengenalan pola biasa dilakukan untuk mengklasifikasi data berdasarkan pengetahuan yang didapatkan dari data yang sudah ada sebelumnya (Pradhan, A., 2012).



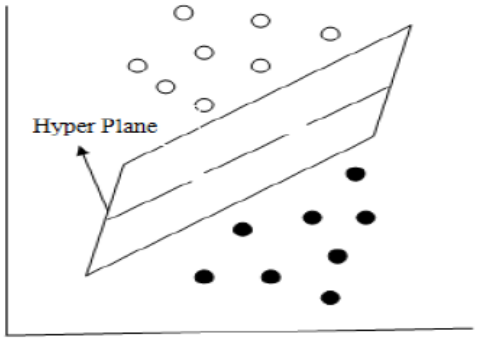
Gambar 3. SVM Model

Sumber: Pradhan, A., 2012. Support vector machine-a survey

Gambar 3 merupakan simple model dari SVM. Model ini berisi dua pola berbeda dan SVM memisahkan dua pola ini. Garis w.x-b=0 adalah garis yang memisahkan dua garis marginal lain. Garis w.x + b = 1 dan w.x + b = -1 adalah garis pada setiap sisi marginal, dimana w adalah *weights*, dimana ini akan menjadi penentu seberapa pengaruh x atau *input* kepada *output* ,x adalah variable input

atau data di positive maupun negative dan b adalah bias yang merupakan batas antara kelas positive dan negative (nilai b=0) bisa diartikan jika formula w.x menghasilkan nilai diatas b maka termasuk dalam positive dan begitu pula sebaliknya .

Ketiga garis ini membentuk *hyper lane* yang memisahkan data positive dan negative dengan data yang terletak di pinggir *hyper lane* disebut *support vector*. Data yang berada di pinggir *hyper lane* ini akan menjadi penentu seberapa tepat *hyper lane* yang akan dihasilkan, sedangkan data yang berada jauh dari *hyper lane* tidak akan menjadi penentu ketepatan *hyper lane.* Proses penentuan *hyperplane* ini juga perlu keseimbangan antara jarak pada kelas positive dan negative. Contoh *hyperplane* ini bisa dilihat pada gambar 4



Gambar 4. *Hyper Plane*

Sumber: Pradhan, A., 2012. Support vector machine-a survey

Tujuan utama dari SVM adalah untuk menemukan *hyper plane* terbaik dari pola input atau biasa disebut dengan *maximum margin.* Hal ini menjadi penyebab SVM biasa disebut juga dengan *Maximum Margin Classifier*. Persamaan yang ditunjukkan di bawah ini adalah representasi dari *hyper plane:*

$$w.x - b = 0$$

dimana;

w = *weight*

x = data

b = bias

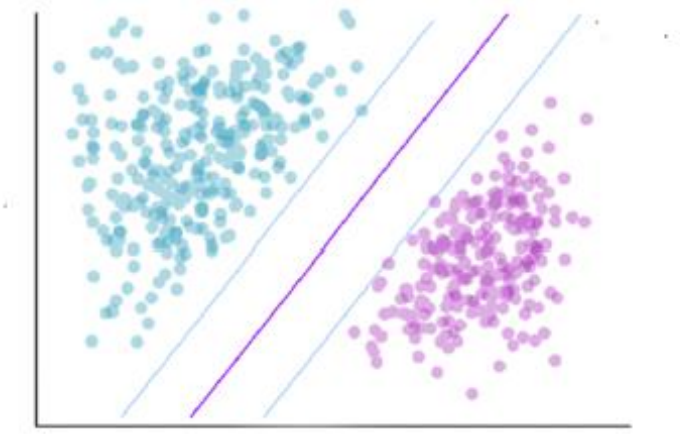
Terdapat beberpa pola garis dalam proses memisahkan data, baik itu garis lurus, melengkung, ataupun ada juga yang perlu ditambahkan dimensi yang lebih tinggi untuk memisahkan data tersebut. Hal ini biasa dilakukan dengan menggunakan kernel trick. Beberapa kernel yang biasa digunakan yakni:

1. Kernel Linear

Fungsi kernel ini biasa digunakan untuk klasifikasi data yang sudah dianalisis dapat dipisah secara linear. Persamaan fungsi kernel linear adalah

$$\tilde{y} = (\tilde{w}^{T}\tilde{x} + b)$$

dimana $\tilde{x}$ data dari label yang ada dalam dataset, jika melihat dari gambar 5 bisa dilihat sebagai titik-titik yang berwarna biru dan ungu, $\tilde{w}$ adalah *weight* dan b adalah bias



Gambar 5. Kernel Linear

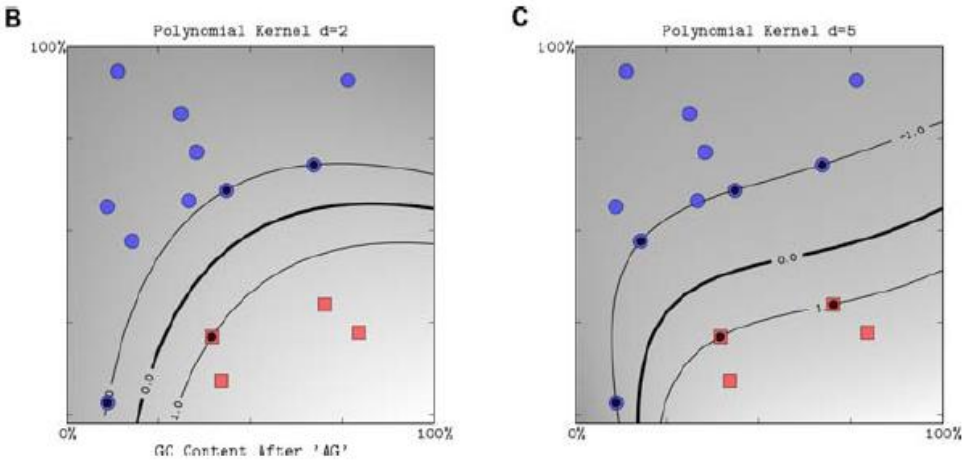
Sumber: Horino, Hiroki (2017). Development of an Entropy-Based Feature Selection Method and Analysis of Online Reviews on Real Estate

2. Kernel Polynomial

Kernel Polynomial adalah bentuk yang lebih umum dari kernel linear. Kernel polinomial tidak hanya melihat fitur yang diberikan dari dataset input yang sudah dinormalisasi untuk menentukan kesamaannya, tetapi juga kombinasinya. Persamaan fungsi kernel polynomial adalah:

$$K(x1, x2) = (x1 \,.\, x2 + c)^d$$

Pada kernel polynomial terdapat parameter derajat (d) dengan nilai default d = 2 jika d = 1 maka itu hanya kernel linear, yang bisa diartikan nilai d harus selalu diatas 1 perbandingan nilai d bisa dilihat pada gambar 2.5-5, sedangkan c adalah bagaimana cara program menyelesaikan masalah atau parameter program memberikan nilai w (*weight*) sehingga akan mempengaruhi seberapa luas margin. Arti dari nilai c ini adalah nilai c yang kecil akan memberikan margin yang lebih luas namun akan terjadi beberapa kesalahan klasifikasi. jika nilai c adalah 0 maka akan membuat *hyperplane* tidak akan mengklasifikasi apapun atau bisa dikatakan nilai b atau bias adalah 0, dan jika nilai c terlalu besar maka membuat program tidak akan mentoleransi adanya nilai b atau bias, yang akan mempengaruhi nilai klasifikasi menjadi lebih kecil disebabkan banyak data yan dianggap *noisy data* oleh program



Gambar 6. Kernel Polynomial

Sumber Ben-Hur, Asa & Ong, Cheng Soon. (2008). Support Vector Machines and Kernels for Computational Biology.

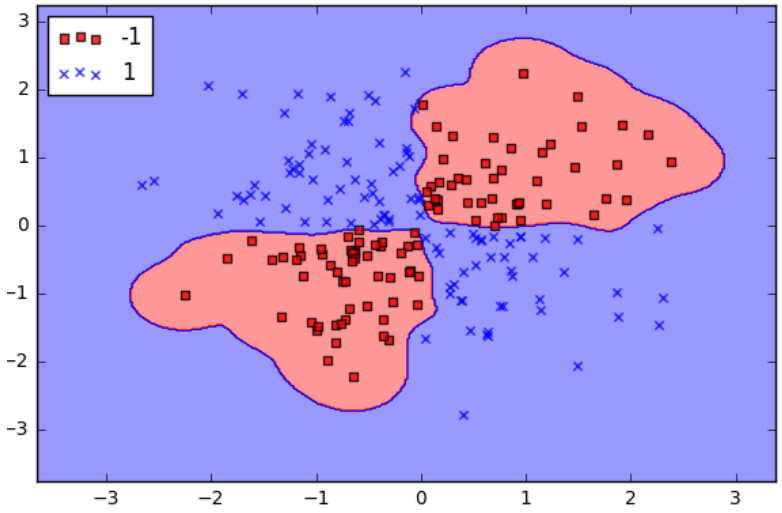
3. Kernel *Radial Basic Function* (RBF)

Kernel RBF atau juga disebut kernel Gaussian adalah konsep kernel yang paling banyak digunakan untuk memecahkan masalah klasifikasi data yang tidak dapat dipisahkan secara linear. Kernel RBF membantu memisahkan data ketika tidak ada pembeda khusus yang diketahui dari setiap data.

Kernel RBF menempatkan fungsi basis radial yang berpusat di setiap titik, kemudian melakukan manipulasi linier untuk memetakan titik ke ruang berdimensi lebih tinggi yang lebih mudah dipisahkan. Persamaan fungsi kernel RBF adalah:

$$K(x, xi) = \exp\left(-\frac{\|x1 - x2\|^2}{2\sigma^2}\right)$$

Dimana σ adalah *variance* dari *hyperparameter*, dan ||x1 – x2|| adalah aturan *Euclidiean* yang berupa jarak dari x1 ke x2. Exp adalah *exponential function* dalam matematika dan memiliki nilai 2.71828183.



Gambar 7. Kernel RBF

Sumber chrisalbon.com/code/machine_learning/support_vector_machines/

2.4.5 *Model Evaluation*

Model evaluation merupakan cara untuk mengevaluasi performa klasifikator. Salah satu cara yang dilakukan adalah dengan menggunakan *confusion matrix. Confusion matrix* merupakan alat yang sangat berguna untuk menganalisis seberapa baik klasifikator mengenali objek dari kelas-kelas yang berbeda. Adapun parameter penting yang didapatkan berdasarkan *confusion matrix* adalah akurasi, presisi, dan *recall*. Nilai *precision* adalah nilai sensitifitas atau nilai ketepatan sistem antara informasi yang diberikan oleh sistem untuk menunjukan secara benar data antara label data. Sedangkan nilai *recall* adalah nilai yang menunjukan tingkat keberhasilan atau spesifisitas untuk mengetahui kembali sebuah informasi

secara benar tentang data pada label. Akurasi adalah nilai yang menunjukan tingkat kedekatan antara nilai prediksi sistem dengan nilai prediksi manusia (Azhari, M., Situmorang, Z., Rosnelly, R., 2021). Rumus mencari nilai akurasi, presisi, dan *recall* yaitu:

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

$$\text{Presisi} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

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

dimana:

- *True Positive* (TP) adalah data dari kelas "A" diprediksi menjadi kelas "A"
- *True Negative* (TN) adalah data kelas lain yang tidak diprediksi menjadi "A" atau sebaliknya
- *False Positive* (FP) adalah data dari kelas lain diprediksi menjadi kelas "A"
- *False Negative* (FN) adalah data dari kelas "A"diprediksi menjadi kelas lain

2.4.6 *Pickle Files*

Pickel file merupakan sebuah modul *library python* untuk menyimpan dan membaca data ke dalam sebuah *file.* Dengan membuat *pickle files* maka data bisa bisa disimpan dalam bentuk *binary file* pada penyimpanan sehingga *file* ini bisa dirubah kembali ke dalam bentuk *python* ketika program *python* yang sudah dibuat dibutuhkan (Kong, Q., Siauw, T. and Bayen, A., 2020).

2.5 Penelitian Terkait

Penelitian terkait merupakan penelitian yang telah dilakukan sebelumnya dan terkait dengan topik yang dibahas. Beberapa penelitian terkait yang digunakan sebagai acuan pada penelitian ini adalah sebagai berikut:

1. Osisanwo F.Y, Akinsola J.E.T, Awodele O, Hinmikaiye J.O, Olakanmi O, dan Akinjobi J (2017), dalam jurnal yang berjudul Supervised Machine Learning Algorithms: Classification and Comparison. Permasalahan yang diselesaikan dalam jurnal ini adalah melakukan perbandingan antara ketujuh algoritma *machine learning; Decision Table, Random Forest (RF) , Naïve Bayes (NB) , Support Vector Machine (SVM), Neural Networks (Perceptron), JRip and Decision Tree (J48) using Waikato Environment for Knowledge Analysis (WEKA)machine learning tool*. Perbandingan dilakukan dengan menggunakan data penyakit diabetes dengan 786 contoh dan delapan atribut. Hasil yang didapatkan dari perbandingan ini adalah SVM sebagai algoritma dengan presisi dan akurasi paling tinggi. Algoritma klasifikasi *Naïve Bayes* dan *Random Forest* menghasilkan tingkat akurasi kedua setelah SVM. Penelitian menunjukkan bahwa waktu yang dibutuhkan untuk membangun model dan presisi (akurasi) merupakan faktor di satu sisi; sedangkan statistik *kappa* dan *Mean Absolute Error (MAE)* adalah faktor lain di sisi lain.
2. Indri Monika, Muhammad Tanzil, dan Sutrisno (2018), dalam jurnal yang berjudul Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak. Penelitian ini bertujuan untuk mengetahui hasil klasifikasi penyimpangan pada tubuh kembang anak

dengan data digunakan pada penelitian sebanyak 90 data yang terbagi menjadi 3 kelas. Kelas penelitian ini mewakilkan 3 jenis penyimpangan tumbuh kembang anak yaitu *Down Syndrome*, *Autisme*, dan *Attention Deficit Hyperactivity Disorder (ADHD)*. Hasil akhir dari penilitan ini menghasilkan rata-rata akurasi tertinggi sebesar 63,11% $\lambda = 10$, C = 1, itermax = 200 dan juga menggunakan kernel polynomial. Perbandingan dari hasil klasifikasi kembang anak dengan bantuan psikolog menunjukkan bahwa sistem menghasilkan akurasi yang kurang baik.

3. Pusphita Anna Octaviani, Yuciana Wilandari, dan Dwi Ispriyanti (2014), dalam jurnal yang berjudul Penerapan Metode Klasifikasi *Support Vector Machine (SVM)* pada data akreditasi sekolah dasar (SD) di Kabupaten Magelang. Data untuk penelitian ini diambil dari Badan Akreditasi Nasional Sekolah/Madrasah dengan membagi data menjadi tiga atribut akreditas yaitu A, B, dan C. Hasil penelitian ini menunjukkan bahwa akurasi prediksi klasifikasi SVM menggunakan fungsi kernel RBF adalah 93,902%. Itu dihitung dari 77 dari 82 SD yang diklasifikasikan dengan benar dengan kelas aslinya.
4. Bety Wulan Sari dan Fadholi Fat Haranto (2019), dalam jurnal yang berjudul Implementasi Support Vector Machine untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom dan Biznet. Pada penelitian ini, peneliti menggunakan teknik *text mining* dengan menerapkan algoritma Support Vector Machine untuk analisis sentimen pengguna twitter terhadap pelayanan Telkom dan Biznet dengan jumlah dataset sebanyak 500 tweet yang berasal dari crawling data twitter, terdapat 250 tweet yang dijadikan

dataset pada masing-masing objek. Sejumlah data tersebut akan dipergunakan untuk data training serta data testing dalam proses pembuatan model menggunakan algoritma Support Vector Machine. Metode yang digunakan untuk pengujian model adalah Confusion Matrix sedangkan K-Fold Cross Validation ditujukan untuk untuk membagi data training dan data testing sesuai lipatan yang digunakan. Hasil pengujian yang diperoleh menggunakan metode K-Fold Cross Validation dan Confusion Matrix pada model yang dibuat menggunakan algoritma Support Vector Machine yang memberikan hasil nilai accuracy 79,6%, precision 76,5%, recall 72,8% , dan F1-score 74,6% untuk Telkom, serta accuracy 83,2%, precision 78,8%, recall 71,6%, dan F1-score 75% untuk Biznet.

5. Imron Sanjaya Girsang, Reyhan Achamd Rizal, dan Sidik Apriyadi Prasetyo (2019) dengan judul Klasifikasi Wajah Menggunakan Support Vector Machine (SVM). Sistem klasifikasi wajah adalah suatu aplikasi yang membuat sebuah mesin dapat mengenali wajah seseorang sesuai dengan citra wajah yang telah ditraining dan disimpan di dalam database mesin tersebut. Klasifikasi wajah sendiri dapat dilakukan dengan berbagai cara, salah satunya adalah menggunakan metode support vector machine (SVM). Penelitian ini dilakukan dengan sampling yang di ambil dalam variasi posisi pada sudut kemiringan subjek (- 90°, -70°, -45°, -25°, -5°) dan (+90°, +70°, +45°, +25°, +5°) dengan ukuran citra 640x480. Sistem klasifikasi wajah didalam penelitian ini dibangun dengan menggunakan metode support vector machine (SVM) dan bahasa pemograman Matlap. Penelitian ini

menghasilkan tingkat true detection 90% dan false detection 10% dari jumlah sampel 200 subjek yang digunakan.

Dari penelitian-penelitian ini hal yang bisa didapatkan adalah SVM merupakan *supervised learning* terbaik untuk melakukan klasifikasi teks, dimana penelitian ini bertujuan untuk mengimplementasikan algoritma SVM untuk mengklasifikasikan judul skripsi mahasiswa Unila berdasarkan subject utama pada *repository* karya akhir mahasiswa yang beralamat pada laman digilib.unila.ac.id.

BAB III
METODE PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian dilakukan di Unit Pelaksana Teknik (UPT) Perpustakaan Universitas Lampung dengan menggunakan *Personal Computer (PC).* Waktu penelitian berlangsung dari bulan Juli 2022 hingga bulan Januari 2023, seperti yang disajikan pada tabel berikut.

Tabel 1. Waktu Penelitian

No	Kegiatan	Jul	Aug	Sep	Oct	Nov	Dec	Jan
1	Study Literature	■						
2	Web Scrapping	■						
3	Data Labelling		■					
4	Data Cleaning		■					
5	Feature Extraction			■				
6	Model Training			■				
7	Model Evaluation				■	■	■	
8	Pengujian Model						■	
9	Save Model Data							■
10	Penyusunan Laporan				■			■

3.2 Alat dan Bahan

Beberapa alat dan bahan yang digunakan dalam penelitian ini pada tabel 2 dan 3

Tabel 2. Alat

No	Alat	Keterangan
1	*Python*	Versi 3.9.11

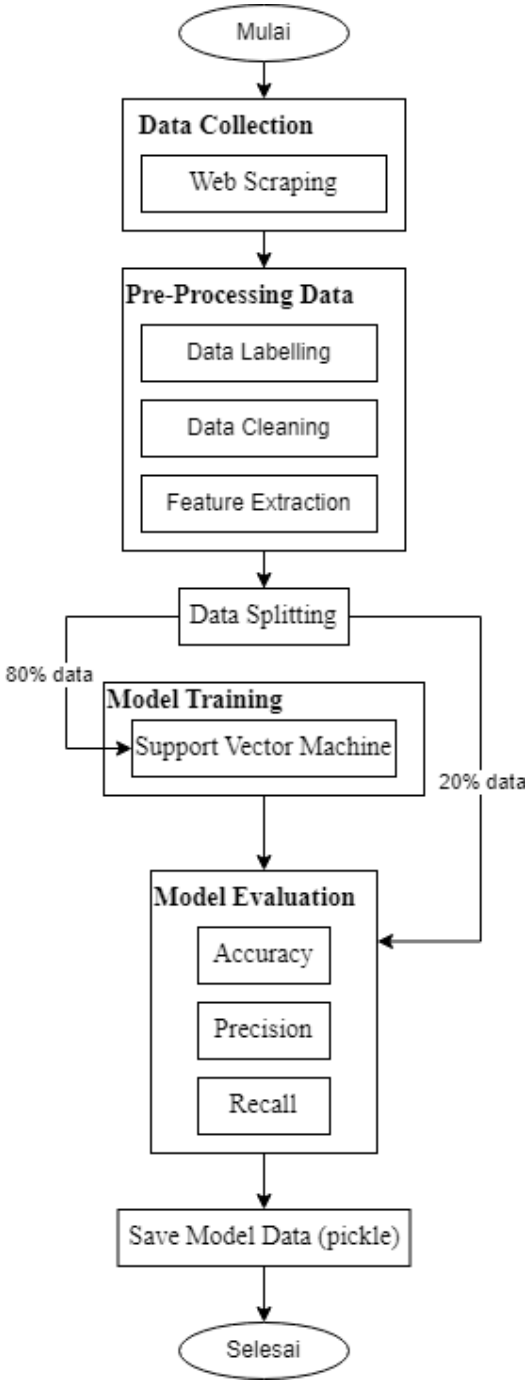
2	*Jupyter Notebook*	Versi 6.4.10	
3	*Python Library*	*sklearn*	*LabelEncoder, TfidfVectorizer, train_test_split, confusion_matrix, SparsePCA,SVC, classification_report, learning_curve*
		pandas	
		numpy	
		nltk	*stopwords*
		joblib	
		matplotlib.pyplot	
		seaborn	
4	Ms. Excel	Ms. Excel 2016	
5	*Add-on*	*DataScraper*	

Tabel 3. Bahan yang digunakan

No	Bahan	Keterangan
1	Data judul skripsi mahasiswa UNILA (*softfile)*	Data judul skripsi mahasiswa UNILA yang tersedia pada website digilib.unila.ac.id pada tanggal 30 Juli 2022 yang berjumlah 1707 data

3.3 Tahapan Penelitian

Konsep metodologi penelitian dilakukan dengan metode *machine learning life cycle*, yang akan digambarkan dengan bagan alur pada gambar 8 berikut:

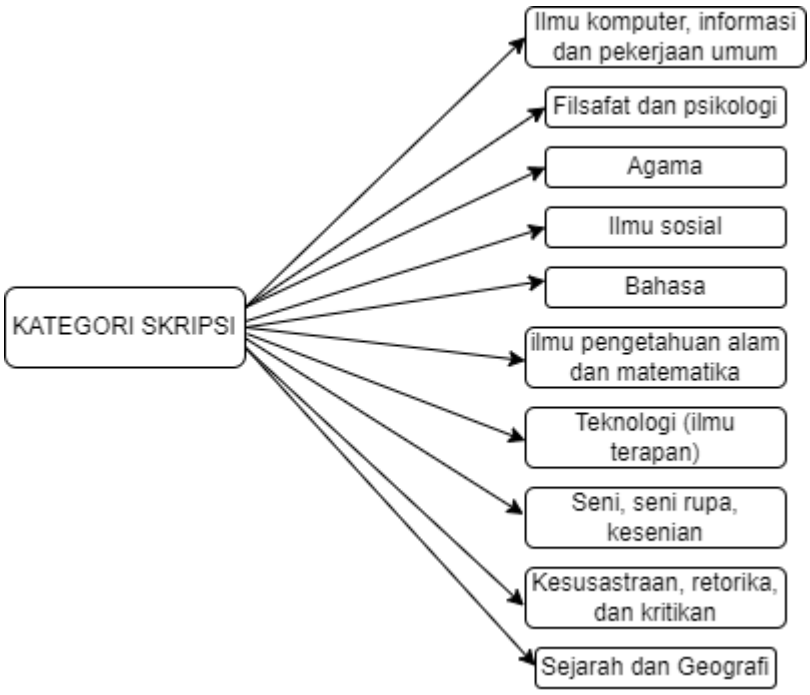


Gambar 8. Bagan Alur Penelitian

Dengan penjabaran metode *machine learning* adalah sebagai berikut:

1) Data Collection

Data Collection dilakukan untuk mengambil dan mengumpulkan data skripsi mahasiswa Unila pada website https://digilib.unila.ac.id/ dengan melakukan *web scrapping*. Proses ini dilakukan dengan bantuan *browser chrome extension* yang bernama *DataMiner*. Data yang dikumpulkan adalah nama, NPM mahasiswa, dan judul skripsi dengan perkiraan jumlah data adalah 1707 data. Pembagian kategori skripsi pada data mahasiswa dilakukan dengan mengikuti pembagian kategori yang ada pada sistem klasifikasi DDC pada *website* digilib.unila.ac.id.



Gambar 9. Pembagian kategori skripsi yang digunakan

2) Pre-processing Data

Tahapan *Pre-processing data* yang dilakukan pada proses ini adalah *data labelling*, *data cleaning* dan *feature extraction.*

a. *Data labelling* dilakukan secara manual untuk menentukan data yang akan diberikan label. Dataset akan dibagi ke dalam empat label yaitu, nama, npm, judul, dan *subject.*

Tabel 4. Deskripsi Data

Atribut	Keterangan
Nama dan NPM	Merupakan sebutan yang diberikan untuk membedakan mahasiswa satu sama lain
Judul	Merupakan nama yang dipakai pada skripsi
Subject	Merupakan kategori yang mengandung informasi dari skripsi yang terdapat dalam koleksi perpustakaan Unila

b. Tahapan selanjutnya adalah *data cleaning* dimana proses ini dilakukan dengan menggunakan *stopword removal,* lalu membuang tanda baca dan angka yang ada pada label judul. Hal ini diperlukan agar data yang diekstrak adalah data bersih.

c. Proses selanjutnya adalah *feature extraction.* Proses ini dilakukan untuk mengubah data yang sudah diberi label menjadi bentuk *vector numerik* dari data skripsi yang telah dilakukan *data cleaning.* Proses *feature extraction* dilakukan dengan menggunakan TF-IDF dari *library scikit-learn*

3) Model Training

Proses ini dimulai dengan melakukan pembagian data atau *data splitting* dengan mengikuti langkah dari Philip Koshute maka jumlah data yang akan digunakan untuk melakukan training model pada penelitian ini adalah 80% dari 1707 data,

sehingga didapatkan 80% data untuk proses *training* dan 20% data untuk proses *testing*.

4) Model Evaluation

Proses ini dilakukan untuk mengevaluasi performa klasifikator dengan menggunakan *confusion matrix*. Adapun parameter penting yang didapatkan berdasarkan *confusion matrix* adalah akurasi, presisi, dan *recall*. Jika nilai akurasi semakin mendekati 100 %, maka performa klasifikator semakin tinggi.

5) ***Hasil***

Penelitian ini diharapkan memiliki hasil akhir berupa model dalam bentuk *pickle file* dengan format file adalah pkl.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis, perancangan, dan pengujian dari sistem prediksi judul skripsi mahasiswa Universitas Lampung dengan menggunakan *machine learning*, maka diperoleh beberapa kesimpulan diantaranya sebagai berikut:

1. Algoritma *Support Vector Machine* dapat digunakan untuk melakukan klasifikasi teks yang berupa judul skripsi.
2. Hasil akhir dari model yang dibuat pada tahapan *model training* menggunakan *data training* memilki akurasi sebesar 0,95, sedangkan pada tahapan *model evaluation* menggunakan *data testing* mendapatkan akurasi sebesar 0,65, presisi sebesar 0,54, dan *recall* sebesar 0,45, sedangkan rata-rata akurasi yang didapat melalui proses pengujian skripsi secara langsung adalah 0,5.
3. Hasil pengujian klasifikasi model SVM dengan kernel RBF pada kategori Ilmu Sosial mendapatkan akurasi sebesar 100%. Hasil ini menunjukkan bahwa akurasi pada kategori ini sangat baik yang dimana hal ini sangat dipengaruhi oleh fakta bahwa kategori Ilmu Sosial memiliki data yang lebih banyak dibandingkan kategori lainnya, sehingga model menjadi lebih familiar dengan kategori Ilmu Sosial.
4. Hasil pengujian klasifikasi model SVM dengan kernel RBF pada kategori Agama hanya mendapatkan akurasi sebesar 0%. Hasil ini menunjukkan bahwa model ini sensitif pada ragam kata dan jumlah data yang ada di setiap

kategori. Hal ini dibuktikan saat proses pengujian banyak judul yang seharusnya masuk ke dalam kategori Agama diprediksi oleh model sebagai kategori Ilmu Sosial.

5. Hasil pengujian klasifikasi model SVM dengan kernel RBF pada kategori Bahasa mendapat akurasi sebesar 80% meskipun jumlah *data training* pada kategori Bahasa tidak sebanyak kategori Ilmu Sosial. Hal ini dikarenakan kategori Bahasa banyak berisi bahasa Inggris, sehingga berbeda dengan kategori lainnya dan model ini tidak menggunakan *stopword* dalam bahasa Inggris, sehingga mempengaruhi model dalam mengklasifikasi judul dalam bahasa Inggris

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, maka peneliti memberikan saran untuk penelitian selanjutnya yaitu;

1. Pada penelitian ini hanya mengkategorikan skripsi berdasarkan 9 kategori utama, sehingga untuk penelitian selanjutnya diharapkan agar mampu mengkategorikan skripsi menggunakan sub kategori dari 9 kategori utama.
2. Pada proses *pre-processing data* gunakan *stopwords* yang bisa mengenali setidaknya dua bahasa, bahasa Indonesia dan bahasa Inggris, karena diantara judul skripsi cukup banyak bahasa yang digunakan selain bahasa Indonesia, lalu tambahkan juga proses *stemming* untuk menghilangkan kata-kata yang memiliki imbuhan dan akhiran, sehingga kembali menjadi kata dasar.
3. Peneliti selanjtunya dapat menggunakan algoritma lain selain SVM untuk klasifikasi *subject* sehingga dapat mengetahui perbandingan tingkat akurasi yang dihasilkan.

DAFTAR PUSTAKA

Ben-Hur, A., Ong, C.S., Sonnenburg, S., Schölkopf, B. and Rätsch, G., 2008. Support vector machines and kernels for computational biology. PLoS computational biology, 4(10), 1000-1173.

Cristianini, N., Shawe-Taylor, J., 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Dalal, M.K. and Zaveri, M.A., 2011. Automatic text classification: a technical review. International Journal of Computer Applications, 28(2), 37-40.

Darujati, C. and Gumelar, A.B., 2012. Pemanfaatan teknik supervised untuk klasifikasi teks bahasa indonesia. Jurnal Bandung Text Mining, 16(1), 5-1.

García, S., Luengo, J. and Herrera, F., Data Preprocessing in Data Mining. Intelligent Systems Reference Library. 2015. doi, 10, 978-3.

Hamakonda, T.P., 1978. Pengantar klasifikasi persepuluhan Dewey. BPK Gunung Mulia.

Han, J., Pei, J., Tong, H., 2022. Data mining: concepts and techniques. Morgan kaufmann.

Horino, H., Nonaka, H., Carreón, E.C.A. and Hiraoka, T., 2017, December. Development of an entropy-based feature selection method and analysis of online reviews on real estate. In 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2351-2355

Kong, Q., Siauw, T. and Bayen, A., 2020. *Python Programming and Numerical Methods: A Guide for Engineers and Scientists*. Academic Press.

Koshute, P., Zook, J. and McCulloh, I., 2021. Recommending Training Set Sizes for Classification. *arXiv preprint arXiv:2102.09382*.

Lavin, M., 2019. Analyzing documents with TF-IDF.

Müller, A.C. and Guido, S., 2016. Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc.

Mitchell, T.M., 1997. Machine learning. McGraw-hill. New York.

Nasteski, V., 2017. An overview of the supervised machine learning methods. Horiz. B 4, 51–62.

Octaviani, P.A., Wilandari, Y. and Ispriyanti, D., 2014. Penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang. Jurnal Gaussian, 3(4), 811-820.

Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138.

Pasaribu, M., Widjaja, A., 2022. Artificial Intelligence: Perspektif Manajemen Strategis. Kepustakaan Populer Gramedia.

Pradhan, A., 2012. Support vector machine-a survey. International Journal of Emerging Technology and Advanced Engineering, 2(8), 82-85.

Turland, M., 2010. Phparchitect's guide to web scraping with PHP. Marco Tabini & Associates, Inc., Toronto, Ont.

Yoo, J.-Y., Yang, D., 2015. Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier. Presented at the Computer and Computing Science 2015, 263–266.

Zaharchuk, G., Gong, E., Wintermark, M., Rubin, D. and Langlotz, C.P., 2018. Deep learning in neuroradiology. American Journal of Neuroradiology, 39(10), 1776-1784.