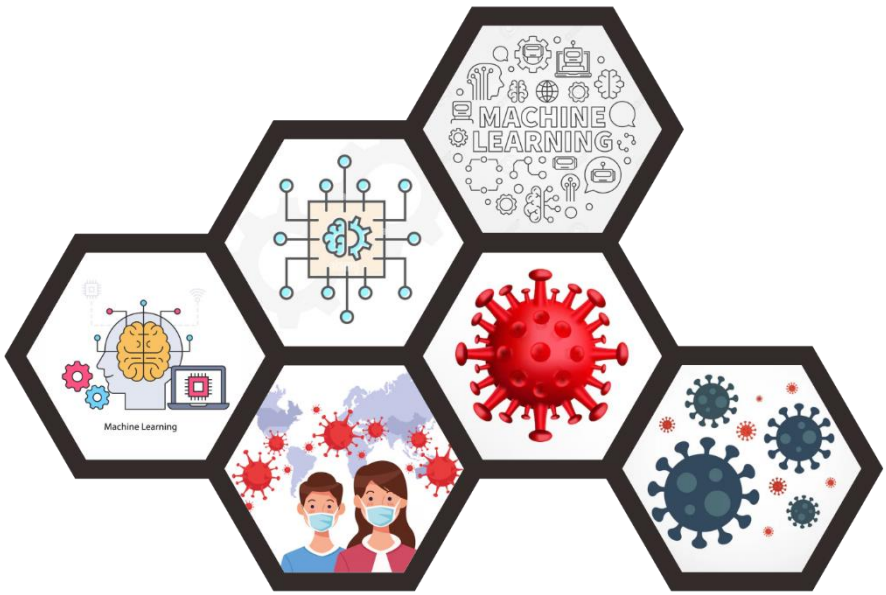
ANALISIS PERBANDINGAN K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE DALAM KLASIFIKASI PASIEN COVID-19 DI KOTA MAKASSAR



MUAMMAR ASHARI ABUSPIN
H062211005

PROGRAM STUDI MAGISTER STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2024

ANALISIS PERBANDINGAN K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE DALAM KLASIFIKASI PASIEN COVID-19 DI KOTA MAKASSAR

MUAMMAR ASHARI ABUSPIN

H062211005

PROGRAM STUDI MAGISTER STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2024

TESIS

ANALISIS PERBANDINGAN K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE DALAM KLASIFIKASI PASIEN COVID-19 DI KOTA MAKASSAR

MUAMMAR ASHARI ABUSPIN
H062211005

telah dipertahankan di depan Panitia Ujian Magister pada 16 Agustus 2024
dan dinyatakan telah memenuhi syarat kelulusan

pada

Program Studi Magister Statistika
Departemen Statisika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Hasanuddin
Makassar

Mengesahkan:

Pembimbing Utama

Dr. Erna Tri Herdiani, S.Si., M.Si.
NIP. 19750429 200003 2 001

Pembimbing Pendamping

Prof Dr. Dr. Georgina Maria Tinungki M.Si.
NIP. 196209261987022001

Ketua Program Studi
Magister Statistika

Dr. Erna Tri Herdiani, S.Si., M.Si.
NIP. 19750429 200003 2 001

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Hasanuddin

Dr. Eng. Amiruddin, M.Si.
NIP. 19720515 199702 1 002

PERNYATAAN KEASLIAN TESIS
DAN PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa, tesis berjudul " Analisis Perbandingan K-Nearest Neighbor Dan Support Vector Machine dalam Klasifikasi Pasien Covid-19 Di Kota Makassar " adalah benar karya saya dengan arahan dari tim pembimbing (Dr. Erna Tri Herdiani, S.Si.,M.Si. dan Prof Dr. Dr Georgina Maria Tinungki M.Si.). Karya Ilmiah ini belum diajukan dan tidak sedang diajukan dalam bentuk apapun kepada perguruan tinggi manapun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka tesis ini. Sebagian dari tesis ini akan dipublikasikan di SCIK sebagai artikel dengan judul " *Comparative Analysis Of K-Nearest Neighbor And Support Vector Machine In Classification Of Covid 19 Disease In Makassar City*".

Dengan ini saya melimpahkan hak cipta dari karya tulis saya berupa tesis ini kepada Universitas Hasanuddin.

Makassar, 16 Agustus 2024

MUAMMAR ASHARI ABUSPIN
NIM. H062211005

UCAPAN TERIMA KASIH

Segala puji hanya milik Allah *Subhanallahu Wa Ta'ala* atas limpahan rahmat dan hidayah-Nya kepada penulis. Shalawat dan salam tercurahkan kepada Rasulullah *Shallallahu 'Alaihi Wa sallam*, keluarganya, *tabi'in*, *tabi'ut tabi'in*, serta orang-orang sholeh yang haq hingga kadar Allah berlaku atas diri mereka. *Alhamdulillaahirobbil'aalamiin*, berkat rahmat dan kemudahan dari Allah *Subhanallahu Wa Ta'ala*, penulis dapat menyelesaikan tesis berjudul " Analisis Perbandingan K-Nearest Neighbor Dan Support Vector Machine dalam Klasifikasi Pasien Covid-19 Di Kota Makassar " sebagai salah satu syarat memperoleh gelar magister pada Program Studi Magister Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

Terima Kasih yang tak terhingga kepada keempat orang tuaku tercinta Ayah **Drs. Jalali Condeng** dan Ibu **Hapidah** yang selalu mendengarkan keluh kesahku, memberikan kasih sayang tak terhingga, materi, semangat, motivasi dan doa yang tak pernah putus, juga kepada saudara-saudaraku, adikku (**Muhammad Aslam Abuspin dan Muhammad Akmal Abuspin**) yang selalu memberikan dukungan, semangat dan dalam penyelesaian tesis ini . Ucapan rasa hormat dan juga terima kasih yang tulus kepada:

1. Yth. **Prof. Dr. Ir. Jamaluddin Jompa, M.Sc** selaku Rektor Universitas Hasanuddin
2. Yth. **Dr. Eng. Amiruddin, M.Si** selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam beserta seluruhnya jajarannya.
3. Yth. **Dr. Anna Islamiyati, S.Si., M.Si** selaku Ketua Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam dan sekaligus sebagai penguji yang telah bersedia menguji serta memberikan masukan-masukan dan arahan dalam penyusunan tesis.
4. Yth. **Dr. Erna Tri Hardiani, S.Si., M.Si** selaku Ketua Program Studi Magister Statistika dan sekaligus pembimbing utama yang senantiasa meluangkan waktu, tenaga, pemikiran dalam membimbing dan mengarahkan penulis dalam menyelesaikan tesis.
5. Yth. **Prof. Dr. Dr Georgina Maria Tinungki, M.Si** sebagai pembimbing yang senantiasa meluangkan waktu, tenaga, pemikiran dalam membimbing dan mengarahkan penulis dalam menyelesaikan tesis.
6. Yth. **Prof.Dr. Nurtiti Sunusi, S.Si., M.Si** selaku penguji yang telah bersedia menguji serta memberikan masukan-masukan dan arahan dalam penyusunan tesis
7. Yth. **Dr. Nirwan,M.Si** selaku penguji yang telah bersedia menguji serta memberikan masukan-masukan dan arahan dalam penyusunan tesis
8. Sahabat terbaik penulis yang bernama **Uddin** yang telah dengan tulus membantu dan mendukung saya untuk terus berjuang menyelesaikan tesis ini. Kehadiranmu menjadi sumber kekuatan yang luar biasa bagi saya, Terima Kasih Terkasih

9. Sahabatku tercinta **Nurul Fadillah, Alfi Nurkhauly, Mita Astuti,** dan **Misrianti** yang memberi semangat bertubi-tubi bagi penulis, si pemberi pelangi setelah hujan.
10. Teman-teman seperjuangan **Dwi Auliyah, Irwan Usman, Maharani, Fadlyansyah Nur Nasruddin, dan Haura.**
11. Teman-teman seperjuangan ruang diskusi yang selalu memberikan semangat dan bantuan
12. Semua pihak yang telah membantu penulis yang tidak bisa disebutkan satu per satu terimakasih atas doa serta dukungannya

Semoga Tuhan Yang Maha Esa memberikan balasan yang berlipat ganda, kasih dan hikmat-Nya atas segala kebaikan yang telah diberikan kepada penulis. Penulis menyadari bahwa masih banyak kekurangan dalam tesis ini, untuk itu dengan segala kerendahan hati, penulis memohon maaf.

Makassar, 16 Agustus 2024

Muammar Ashari Abuspin

ABSTRAK

MUAMMAR ASHARI ABUSPIN. **Analisis Perbandingan K-Nearest Neighbor Dan Support Vector Machine dalam Klasifikasi Pasien Covid-19 Di Kota Makassar** (dibimbing oleh Dr. Erna Tri Herdiani, S.Si.,M.Si. dan Prof Dr. Dr Georgina Maria Tinungki M.Si.)

Latar Belakang. Penyakit virus corona 2019 atau yang lebih dikenal dengan COVID-19 merupakan wabah yang pertama kali terdeteksi di kota Wuhan, Tiongkok pada bulan Desember 2019. Sebelum disebut COVID-19, WHO atau Organisasi Kesehatan Dunia memberi nama sementara pada virus baru ini sebagai Virus Corona Baru 2019 (2019). ncov). Dan pada tanggal 21 April 2020 WHO resmi menyebut virus 2019-nCoV sebagai COVID-19. Ada 4 faktor yang mempengaruhi pasien COVID 19 dan faktor-faktor ini akan dipertimbangkan. Untuk menganalisis pengaruh faktor-faktor tersebut, algoritma K-Nearest Neighbor (kNN) dan Support Vector Machine (SVM) menggunakan JASP. **Tujuan** dari penelitian ini adalah untuk mengetahui perbandingan tingkat akurasi klasifikasi K-Nearest Neighbor dan Support Vector Machine terhadap pasien Covid 19. **Hasil** penelitian menunjukkan bahwa SVM mencapai tingkat akurasi yang lebih tinggi yaitu 98,43% dibandingkan metode kNN yang menghasilkan akurasi sebesar 98,40%, jika diterapkan pada data pasien COVID 19 di kota Makassar.

Kata Kunci:Covid 19 ; K Nearest Neighbor ; Support Vector Machine.

ABSTRACT

MUAMMAR ASHARI ABUSPIN. **Comparative Analysis of K-Nearest Neighbor and Support Vector Machine in Classifying Covid-19 Patients in Makassar City** (supervised by Dr. Erna Tri Herdiani, S.Si.,M.Si. dan Prof Dr. Dr Georgina Maria Tinungki M.Si.).

Background. The 2019 corona virus disease or better known as COVID-19 is an outbreak that was first detected in the city of Wuhan, China in December 2019. Before it was called COVID-19, WHO or the World Health Organization gave this new virus a temporary name as the New Corona Virus 2019 (2019). ncov). And on April 21 2020 WHO officially called the 2019-nCoV virus COVID-19. There are 4 factors that influence COVID 19 patients and these factors will be considered. To analyze the influence of these factors, the K-Nearest Neighbor (kNN) algorithm and Support Vector Machine (SVM) use JASP. **Objective** The purpose of this research is to find out the comparison of the accuracy levels of K-Nearest Neighbor and Support Vector Machine classification for Covid 19 patients. **Results** Research shows that SVM achieves a higher level of accuracy, namely 98.43% compared to the kNN method which produces an accuracy of 98.40%, when applied to COVID 19 patient data in the city of Makassar.

Keywords: Covid 19 ; K Nearest Neighbor ; Support Vector Machine.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1. Latar Belakang

Penyakit Coronavirus 2019 atau lebih dikenal dengan istilah COVID-19 merupakan suatu wabah yang awalnya terdeteksi di Kota Wuhan, Cina pada Desember 2019. Sebelum disebut sebagai COVID-19, WHO atau World Health Organization memberikan nama sementara virus baru ini sebagai Coronavirus Novel 2019 (2019-nCoV). Dan pada 21 April 2020 WHO secara resmi menyebut virus 2019-nCoV menjadi COVID-19 (Sohrabi dkk., 2020). COVID-19 bermula dari betacoronavirus (SARS-CoV-2) yang menyerang bagian saluran pernapasan bagian bawah yang berubah menjadi pneumonia di tubuh manusia. Virus COVID-19 merupakan coronavirus jenis baru. COVID-19 dianggap sebagai kerabat dari Severe Acute Respiratory Syndrome (SARS) dan Middle East Respiratory Syndrome Coronavirus (MERS)(Sohrabi dkk., 2020).

Berdasarkan data yang didapat dari WHO, terdapat 179 negara yang sudah terpapar virus COVID-19. Hal menandakan bahwa virus ini memiliki tingkat paparan yang sangat tinggi dan cepat. Cara penyebarannya juga sangat sederhana. Penyebarannya dapat berupa bersin, batuk, atau berinteraksi dengan orang yang sudah terinfeksi. Dan virus ini lebih rentan terhadap orang tua dan mereka yang memang sudah memiliki riwayat penyakit serius. Ada beberapa faktor yang mempengaruhi cepatnya penyebaran virus ini yaitu umur tua, banyaknya orang bepergian ke negara yang sudah terinfeksi, melakukan kontak dengan orang yang terinfeksi, adanya komorbid dan sebagainya (Rasmussen dkk., 2020). Faktor-faktor tersebut dapat menjadi data dan dapat diolah dengan data mining.

Data mining merupakan proses pengumpulan dan pengolahan data yang bertujuan untuk mengekstrak informasi penting yang terdapat pada data. Dalam dunia kesehatan penggunaan metode data mining telah banyak membantu dunia kesehatan dalam membuat prediksi mengenai masalah kesehatan yang dihadapi. Salah satunya penyakit Covid 19, Penyakit Covid 19 ini dapat menyebabkan kematiaan.

Ada tiga metode dari data mining yaitu, prediction, Association, dan Segmentation. Tipe Prediction terbelah menjadi tiga yaitu Classification, Regression, dan Time Series. Classification menggunakan algoritma diantaranya Decision Trees, Neural Networks, Support Vector Machine, kNN, Naïve Bayes, dan GA. Dalam melakukan metode Klasifikasi, ada proses estimasi yang bernama simple/single split yaitu memisahkan data untuk training (70%) dan testing (30%). Hal ini digunakan untuk melihat prediksi metode klasifikasi tersebut. Klasifikasi bekerja melalui pengenalan pola atau model dari sebuah kelas. Klasifikasi bertujuan agar pola tersebut bisa dipakai

dalam melakukan prediksi ataupun klasifikasi pada seseorang yang didasarkan pada analisis data latih (Nurmasani & Pristyanto, 2021).

Klasifikasi digunakan untuk mengelompokkan data kedalam beberapa kategori agar lebih mudah untuk diolah dan dianalisis (Devita dkk., 2018). Metode klasifikasi yang umum digunakan pada disiplin ilmu statistika adalah Analisis Diskriminan dan Regresi Logistik. Namun, semakin populernya era data yang menunjukkan bahwa terjadinya pertumbuhan pesat dari volume data yang luar biasa banyak sehingga menghasilkan set data besar, maka sangat dibutuhkan alat analisis yang kuat dan serbaguna untuk mengungkap informasi berharga dari set data besar dan untuk mengubah set data besar tersebut menjadi pengetahuan yang terorganisir (J. Han & Pei, 2012).

Klasifikasi merupakan salah satu teknik dalam *text mining* dan *data mining* yang digunakan dalam proses mencari model atau fungsi yang menjelaskan atau membedakan kelas kelas pada data dan konsep yang bertujuan untuk menggunakan model tersebut dalam melakukan prediksi terhadap *data testing* (Fauzan dkk., 2018). Algoritma Klasifikasi memiliki keunggulan dan kelemahannya masing masing dalam mengklasifikasikan data dalam bentuk teks, diantaranya klasifikasi menggunakan algoritma Support Vector Machine dan kNN memiliki tingkat akurasi tertinggi di bandingkan dengan algoritma yang lainnya. Kelebihan dari algoritma Support Vector Machine (SVM) ialah memiliki akurasi yang tinggi, efisien dalam menggunakan memori dan dapat menangani data yang tidak terdistribusi secara normal. Sedangkan kelebihan algoritma K-Nearest Neighbor (kNN) adalah ketangguhan terhadap data training yang memiliki banyak noise dan data dalam jumlah yang besar.

Penelitian terdahulu yang menggunakan dataset Data Science for COVID-19 (DS4C) yang diambil dari kaggle yang juga digunakan pada penelitian ini pernah dilakukan oleh Al-Najjar dan Al-Rousan membahas mengenai prediksi kesembuhan dan kematian pasien Covid-19 di Korea Selatan dengan algoritma yang digunakan yaitu Artificial Neural Network (ANN) (Al-Najjar & Al-Rousan, 2020).

Penelitian dengan dataset yang sama dilakukan juga oleh Muhammad dkk. yang membahas tentang penggunaan beberapa model untuk mendapatkan akurasi tertinggi. Model yang digunakan antara lain Decision Tree, Support Vector Machine, Naïve Bayes, Logistic Regression, Random Forest, dan K-Newarest Neighbor. Dengan akurasi paling tinggi yang didapatkan dari beberapa model yang digunakan adalah Decision Tree memiliki akurasi yang tertinggi dengan akurasi 99.85% (Muhammad dkk., 2020). Kekurangan dari penelitian ini adalah terpaku terhadap akurasi tanpa melihat matriks yang mempengaruhi baik atau buruknya sebuah model yang dibuat. Melihat dari penelitian sebelumnya, penelitian yang akan dilakukan kali

ini adalah melakukan klasifikasi dengan atribut yang digunakan yaitu Jenis Kelamin, Umur, serta Komorbid dengan menggunakan model klasifikasi Support Vector Machine, dan K-Nearest Neighbor dengan tujuan membandingkan tingkat akurasi dari ke dua algoritma tersebut, maka peneliti memutuskan untuk menggunakan metode metode K- Nearest Neighbor dan SVM. Kinerja dari kedua metode tersebut akan dibandingkan sehingga dapat diketahui metode yang paling efektif dalam melakukan klasifikasi. Berdasarkan pada latar belakang diatas maka penulis ingin melakukan penelitian dengan judul "ANALISIS PERBANDINGAN K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE DALAM KLASIFIKASI PASIEN COVID 19 DI KOTA MAKASSAR"

1.2. Rumusan Masalah

Berdasarkan latar belakang yang diperoleh, maka didapatkan rumusan masalah sebagai berikut:

1. Bagaimana tingkat akurasi klasifikasi *K-Nearest Neighbor* terhadap pasien Covid 19 ?
2. Bagaimana tingkat akurasi klasifikasi *Support Vector Machine* terhadap pasien Covid 19 ?
3. Bagaimana perbandingan tingkat akurasi klasifikasi *K-Nearest Neighbor* dan *Support Vector Machine* terhadap pasien Covid 19 ?

1.3. Tujuan Penelitian

Tujuan yang ingin dicapai pada penelitian ini adalah sebagai berikut:

1. Menentukan tingkat akurasi klasifikasi *K-Nearest Neighbor* terhadap faktor-faktor pasien Covid 19
2. Menentukan tingkat akurasi klasifikasi *Support Vector Machine* terhadap faktor-faktor pasien Covid 19
3. Menentukan perbandingan tingkat akurasi klasifikasi *K-Nearest Neighbor* dan *Support Vector Machine* terhadap pasien Covid 19

1.4. Batasan Masalah

Dalam penelitian ini permasalahan dibatasi pada perbandingan hasil keakuratan klasifikasi menggunakan metode *K-Nearest Neighbor* dan *Support Vector Machine*.

1.5. Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Menambah wawasan dan pengetahuan cara menentukan perbandingan tingkat akurasi klasifikasi *K-Nearest Neighbor* dan *Support Vector Machine* terhadap pasien Covid 19.

2. Dapat dijadikan sebagai salah satu rujukan bagi pemerintah agar lebih memperhatikan penderita pasien Covid 19 di Indonesia secara umum dan di kota Makassar secara khusus.

1.6. Kajian Teori

1.6.1. Data Mining

Menurut (Kana dkk., 2022), Data mining adalah proses yang memperkerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis. Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Rivandi dkk., 2019). Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Handoko & Lesmana, 2018).

Berdasarkan definisi-definisi di atas tentang Data mining dapat disimpulkan bahwa data mining adalah sebuah proses pencarian secara otomatis untuk menemukan pola atau model dari suatu database yang besar.

Operasi *data mining* menurut sifatnya dibedakan menjadi 2, yaitu bersifat (1) prediksi (*prediction driven*) untuk menjawab pertanyaan apa dan sesuatu yang bersifat abstrak atau transparan. Operasi prediksi digunakan untuk validasi *hipotesis*, *querying* dan pelaporan. (2) penemuan (*discovery driven*) bersifat transparan dan untuk menjawab pertanyaan "mengapa?". Operasi penemuan digunakan untuk analisis data eksplorasi, pemodelan prediktif, segmentasi *database*, analisis keterkaitan (*link analysis*) dan deteksi deviasi (Syahra dkk., 2019).

Beberapa teknik dan sifat *data mining* adalah sebagai berikut:

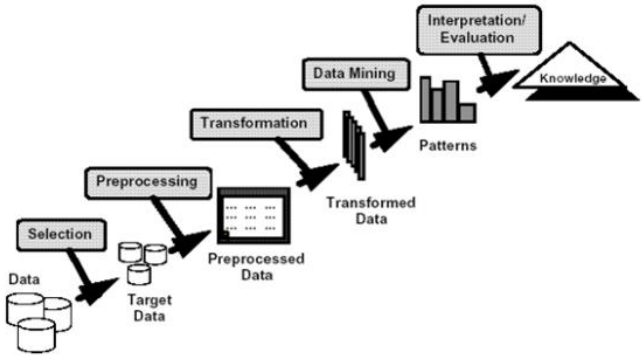
1. Klasterisasi. Adalah mempartisi *data-set* menjadi beberapa *sub-net* atau kelompok sedemikian rupa sehingga elemen-elemen dari suatu kelompok tertentu memiliki *set property* yang di *share* bersama, dengan tingkat similaritas yang tinggi dalam suatu kelompok yang rendah. Disebut juga dengan "*unsupervised learning*".
2. Regresi. Adalah memprediksi nilai dari suatu variabel kontinyu yang diberikan berdasarkan nilai adari variabel yang lain, dengan mengasumsikan sebuah model ketergantungan linier atau nonlinier.
3. Klasifikasi. Adalah menentukan sebuah *record* data baru ke salah satu dari beberapa kategori (kelas) yang telah didefinisikan sebelumnya dan disebut juga dengan "*supervised learning*"

4. Kaidah Asosiasi (*association rule*). Adalah mendeteksi kumpulan atribut-atribut yang muncul bersamaan (*co-occur*) dalam frekuensi yang sering dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut (Informatika & Industri, 2004).

Prediksi/*forcasting* adalah menetukan jumlah kebutuhan bulan mendatang terkait dengan dukungan data historis (*historical data*) atau serangkaian waktu/periode yang dianalisis sehingga dapat diperhitungkan untuk memprediksi jumlah kebutuhan pada bulan mendatang. Prediksi juga dapat digunakan dalam pengklasifikasian, tidak hanya untuk memprediksi *time series*, karena sifatnya yang bisa menghasilkan *class* berdasarkan atribut yang ada (Saputra, 2021).

Knowledge Discovery in Database (KDD) adalah proses menentukan informasi yang berguna serta pola-pola yang ada dalam data. Informasi ini terkandung dalam basis data yang berukuran besar yang sebelumnya tidak diketahui dan potensial bermanfaat. *Data Mining* merupakan salah satu langkah dari serangkaian proses iterative KDD (Hutasuhut dkk., 2019). Berikut tahapan proses KDD dapat dilihat pada gambar 1.

Gambar 1. Tahapan Proses KDD



Tahapan proses KDD terdiri dari:

1. Data *Selection*. Pada proses ini dilakukan pemilihan himpunan data, menciptakan himpunan data target, atau memfokuskan pada subset *variable* (sampel data) dimana penemuan (*discovery*) akan dilakukan. Hasil seleksi disimpan dalam suatu berkas yang terpisah dari basis data operasional.
2. Pre-*Processing* dan *Cleaning* Data. Pre-*Processing* dan *Cleaning* Data dilakukan membuang data yang tidak konsisten dan *noise*, duplikasi data, memperbaiki kesalahan data, dan bisa diperkaya dengan data *eksternal* yang relevan.

3. *Transformation*. Proses ini mentransformasikan atau menggabungkan data ke dalam yang lebih tepat untuk melakukan proses *mining* dengan cara melakukan peringkasan (*agregasi*).
4. *Data Mining*. Proses *Data Mining* yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik, metode atau algoritma tertentu sesuai dengan tujuan dari proses KDD secara keseluruhan.
5. *Interpretation*/Evaluasi. Proses untuk menerjemahkan pola-pola yang dihasilkan dari *Data Mining*. Mengevaluasi (menguji) apakah pola atau informasi yang ditemukan bersesuaian atau bertentangan dengan fakta atau hipotesa sebelumnya. Pengetahuan yang diperoleh dari pola-pola yang terbentuk dipresentasikan dalam bentuk visualisasi.

1.6.2. Klasifikasi

Klasifikasi merupakan kata serapan dari bahasa Belanda, *classificatie*, yang sendirinya berasal dari bahasa Prancis *classification*. Istilah ini menunjuk kepada sebuah metode untuk menyusun data secara sistematis atau menurut beberapa aturan atau kaidah yang telah ditetapkan. Di dalam KBBI, klasifikasi adalah penyusunan bersistem dalam kelompok atau golongan menurut kaidah atau standar yang ditetapkan. Secara harfiah bisa pula dikatakan bahwa klasifikasi adalah pembagian sesuatu menurut kelas-kelas. Menurut Ilmu Pengetahuan, Klasifikasi adalah Proses pengelompokkan benda berdasarkan ciri-ciri persamaan dan perbedaan.

Dalam statistika, klasifikasi adalah masalah untuk mengidentifikasi yang mana dari kumpulan kategori (sub-populasi) yang menjadi observasi baru, berdasarkan kumpulan data pelatihan yang berisi observasi (atau contoh) yang keanggotaan kategorinya diketahui. Contohnya adalah menetapkan email tertentu ke kelas "spam" atau "non-spam", dan menetapkan diagnosis untuk pasien tertentu berdasarkan karakteristik pasien yang diamati (jenis kelamin, tekanan darah, ada atau tidak adanya gejala tertentu, dll.) . Klasifikasi adalah contoh pengenalan pola.

Dalam terminologi pembelajaran mesin klasifikasi dianggap sebagai contoh pembelajaran yang diawasi, yaitu pembelajaran di mana serangkaian pelatihan observasi yang diidentifikasi dengan benar tersedia. Prosedur tanpa pengawasan yang sesuai dikenal sebagai pengelompokan, dan melibatkan pengelompokan data ke dalam kategori berdasarkan beberapa ukuran kesamaan atau jarak yang melekat.

Classification merupakan salah satu dari tiga bagian prediction, sedangkan prediction sendiri merupakan metode dari data mining yang di gunakan dalam proses penarikan data yang sangat besar untuk di terjemahkan kedalam data base yang besar sehingga memudahkan pengambilan keputusan suatu masalah dan juga sebagai prediksi masa

depan. Data mining sendiri mengumpulkan beberapa teknik untuk menemukan pola yang tidak diketahui sebelumnya.

Ada beberapa Klasifikasi yang menggunakan algoritma diantaranya :

1. *Decision Trees*, Algoritma machine learning yang menggunakan seperangkat aturan untuk membuat keputusan dengan struktur seperti pohon yang memodelkan kemungkinan hasil, biaya sumber daya, utilitas dan kemungkinan konsekuensi atau resiko. Konsepnya adalah dengan cara menyajikan algoritma dengan pernyataan bersyarat, yang meliputi cabang untuk mewakili langkah-langkah pengambilan keputusan yang dapat mengarah pada hasil yang menguntungkan.

2. Neural Networks, Algoritma Neural Network merupakan metode yang terinspirasi dari jaringan syaraf otak manusia karena di desain mengikuti cara otak manusia melakukan proses dan menyimpan suatu informasi (Annisa dkk., 2020). Algoritma Neural Network digunakan sebagai tools yang menggambarkan data statistik yang non-linear, dengan menggambarkan suatu hubungan yang kompleks antara input dan output (Singh & Chauhan, 2009a). Neural Network terdiri dari beberapa layer, disetiap layer biasanya terdapat minimal 1 atau lebih Processing Elements (PE). Processing Elements digunakan untuk mensimulasikan cara kerja neuron yang ada didalam otak manusia. Oleh karena itu PE juga sering disebut neuron atau node, setiap PE menerima input dari lapisan sebelumnya (Singh & Chauhan, 2009b).

3. *Support Vector Machine*, Machine Learning merupakan salah satu hal yang berkaitan erat dengan ilmu Data Science. Machine Learning sendiri merupakan bagian dari Artificial Intelligence (AI) yang digunakan untuk meniru hingga menggantikan cara atau perilaku manusia dalam menghadapi dan menyelesaikan permasalahan. Dengan kata lain, Machine Learning adalah mesin yang dilatih secara terus menerus agar dapat mengenal lingkungannya sehingga dapat memiliki pola pikir layaknya manusia dalam pengambilan keputusan. Cara kerja dari metode *Support Vector Machine* khususnya pada masalah non-linear adalah dengan memasukkan konsep kernel ke dalam ruang berdimensi tinggi. Tujuannya adalah untuk mencari hyperplane atau pemisah yang dapat memaksimalkan jarak (margin) antar kelas data. Untuk menemukan hyperplane terbaik, kita dapat mengukur margin kemudian mencari titik maksimalnya. Proses pencarian hyperplane yang terbaik adalah dari metode *Support Vector Machine* ini.

4. kNN, K-Nearest Neighbor, Algoritma kNN atau K-Nearest Neighbor merupakan algoritma klasifikasi yang bekerja dengan mengambil sejumlah K data terdekat (tetangganya) sebagai acuan untuk menentukan kelas dari data baru. Algoritma ini mengklasifikasikan data berdasarkan *similarity* atau

kemiripan atau kedekatannya terhadap data lainnya.Algoritma kNN bekerja dengan menggunakan semua kumpulan data untuk menemukan k-titik terdekat ke sampel baru atau jumlah k sampel yang paling mirip. Algoritma ini biasa digunakan untuk masalah klasifikasi dengan nilai K ditentukan oleh si peneliti. Dalam K-Nearest Neighbor, data point yang berada berdekatan disebut “neighbor” atau “tetangga”. Titik yang memiliki jarak paling dekat akan diklasifikasikan. Ukuran jarak yang digunakan adalah jarak euclidean dan jarak hamming.

5. Naïve Bayes, Naive bayes adalah suatu pengklasifikasian probabilistik sederhana yang melakukan perhitungan terhadap sekumpulan probabilistik dengan penjumlahan frekuensi dan gabungan nilai dari data set (Huda dkk., 2020). *Naive bayes* memerlukan data pelatihan untuk mengestimasi parameter yang dibutuhkan untuk klasifikasi. Selain itu, *naive bayes* mampu menangani nilai yang hilang dengan mengabaikan atribut selama perhitungan estimasi peluang. Kemudian hasil dari model klasifikasi tersebut akan dibandingkan (Mayadewi & Rosely, 2015).

6. GA (Genetic Algorithm), Algoritma Genetika adalah salah satu algoritma yang digunakan untuk mengoptimasi hasil akhir berdasarkan sebaran inputan data acak. Contoh kasus yang akan dibahas kali ini adalah untuk mengoptimasi performa mobil dengan mengupgrade parts tertentu. Algoritma Genetika adalah proses pencarian yang didasarkan pada seleksi alam. Teknik ini secara umum digunakan untuk menghasilkan solusi optimasi dan teknik pencarian. Algoritma Genetika menggunakan teknik yang diinspirasi dari teori evolusi alam, seperti seleksi, warisan, crossover, dan mutasi.

1.6.3. Support Vector Machine

Support Vector Machine (SVM) pertama kali dikenalkan oleh Vapnik pada tahun 1992 sebagai salah satu metode *machine learning* yang bekerja dengan prinsip *Structural Risk Minimization*/SRM yang bertujuan untuk menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space*. Metode ini menggunakan hipotesis berupa fungsi linier dalam sebuah ruang fitur yang berdimensi tinggi, dengan mengimplementasikan *learning* bisa yang berasal dari teori pembelajaran statistik (Nugroho dkk., 2003a).

Support Vector Machine (SVM) adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi (Salma dkk., 2018). *Support Vector Machine* (SVM) memiliki prinsip dasar *linier classifier* yaitu kasus klasifikasi yang secara linier dapat dipisahkan, namun *Support Vector Machine* (SVM) telah dikembangkan agar dapat bekerja pada *problem non-linier* dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi.

Pada ruang berdimensi tinggi, akan dicari *hyperplane* yang dapat memaksimalkan jarak (*margin*) antara kelas data.

Mesin vektor pendukung (*Support Vector Machine* (SVM)) merupakan suatu teknik untuk melakukan prediksi, baik prediksi dalam kasus regresi maupun klasifikasi (Fachrurrazi & Burhanuddin, 2018a). Teknik SVM digunakan untuk mendapatkan fungsi pemisah (*hyperplane*) yang optimal untuk memisahkan observasi yang memiliki nilai variabel target yang berbeda. Metode *Support Vector Machine* memiliki beberapa keuntungan yaitu:

- Generalisasi

Generalisasi didefinisikan sebagai kemampuan suatu metode untuk mengklasifikasi suatu *pattern* atau pola, yang tidak termasuk data yang digunakan dalam fase pembelajaran metode itu.

- *Curse of dimensionality*

Curse of dimensionality didefinisikan sebagai masalah yang dihadapi suatu metode *pattern recognition* dalam mengestimasikan parameter dikarenakan jumlah sampel data yang relatif lebih sedikit dibandingkan dengan dimensional ruang vektor tersebut.

- *Feasibility*

SVM dapat diimplementasikan relatif lebih mudah, karena proses penentuan *support vector* dapat dirumuskan dalam *Quadratic Programing* (QP) problem (Nugroho dkk., 2003b)

Menurut penelitian (Ervinna dkk., 2013) *Support Vector Machine* (SVM) adalah suatu metode atau algoritma untuk melakukan klasifikasi maupun prediksi. Prinsip kerja dari metode ini adalah mencari ruang pemisah yang paling optimal dari suatu dataset dalam kelas yang berbeda. Dalam kehidupan sehari-hari, kita sering diperhadapkan pada persoalan-persoalan yang tidak linear atau data yang tidak dapat benar-benar dipisahkan secara linear yaitu suatu kondisi dimana tidak ada sebuah garis atau bidang yang dapat dibuat untuk menjadi pemisah antar kelas data.

$$f(x_d) = \sum_{i=1}^{ns} a_i y_i \vec{x}_i \vec{x}_d + b \qquad (2.1)$$

Dimana:

ns = Jumlah support vector

αi = Nilai bobot setiap titik data

yi = Kelas data

$\vec{x}\, i$ = Variabel support vector

$\vec{x}\, d$ = Data yang akan diklasifikasikan

b = Nilai error atau bias

Bentuk umum Support Vector Machine (SVM)

Mesin Vektor Pendukung atau *Support Vector Machine (SVM)* menggunakan model linear sebagai *decision boundary* dengan bentuk umum sbb:

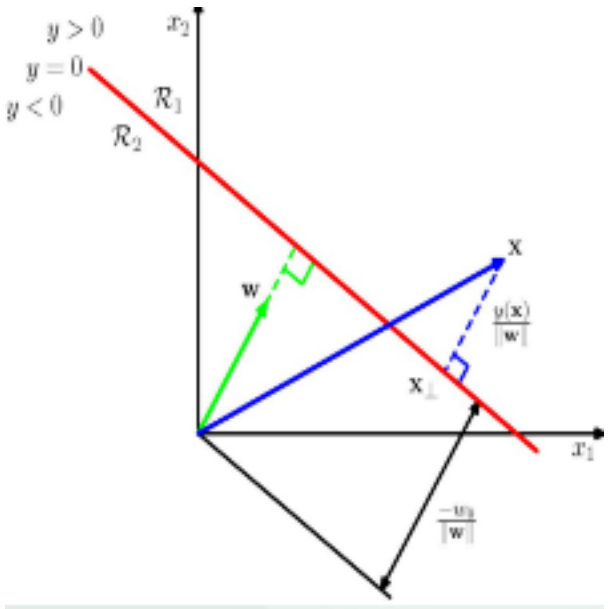
$$y(x) = w^t \emptyset(x) + B \qquad (2.2)$$

dimana x adalah vektor input, w adalah parameter bobot, $\emptyset(x)$adalah fungsi basis, dan B adalah suatu bias (Fachrurrazi & Burhanuddin, 2018b).

Hyperplane

- Bentuk model linear yang paling sederhana untuk decision boundary adalah:
  $$\boldsymbol{y(x) = w^t x + w_0} \qquad (2.3)$$
  Dimana **x** adalah vektor input, **w** adalah vektor bobot dan $\boldsymbol{w_0}$ adalah bias.
- Sehingga, decision boundary adalah y(x)=0, yaitu suatu *hyperplane* berdimensi (D-1)
- Suatu vektor input x akan diklasifikasikan ke kelas 1 (R_1) jika y(x)≥0, dan kelas 2 (R_2) jika y(x)<0

Gambar 2. Hyperplane



Sifat-Sifat Hyperplane

- Jika $\boldsymbol{x_A}$ dan $\boldsymbol{x_B}$ terletak pada *decision boundary* (DS), maka y(x_A)=y(x_B)=0 atau W^T $(x_A - x_B)$ = 0, sehingga **w** tegak lurus terhadap semua vektor di DS. Dengan kata lain **w** menentukan orientasi dari DS
- Jarak titik awal ke DS adalah $-w_0/||\boldsymbol{W}||$. Dengan kata lain w_0 menentukan lokasi DS.
- Jarak sembarang vektor x ke DS dan searah **w** adalah **y(x)/||w||**

Multi-Class Classification

Muti-class classification adalah masalah klasifikasi yang memiliki jumlah kelas lebih dari 2. Sementara SVM standar didesain untuk masalah *twoclass classification*. Ada beberapa teknik yang memungkinkan penggunaan SVM *standar two-class* untuk masalah multi-class (Fachrurrazi & Burhanuddin, 2018b), misal K kelas, yaitu:

• *One-vs-The Rest,* yaitu membangun *K* buah SVM, dimana model ke-k, yaitu y k (x), dilatih dengan menggunakan data dari kelas Ck sebagai sampel positip (+1) dan data dari kelas yang lain sebagai sampel negatip (-1). Contoh: $y_2(x)$ akan memisahkan antara kelas 0 dan kelas -kelas lainnya (1,2,3,4)

• One-vs-One, yaitu membangun *K(K-1)/2* buah SVM yang merupakan semua kemungkinan pasangan kelas, selanjutnya suatu data penguji akan diklasifikasikan ke kelas yang menang paling banyak. Contoh: y_2(x) akan memisahkan kelas 0 dan kelas 1, y_2(x) akan memisahkan kelas 1 dan kelas 2, dst.

Selanjutnya, b dapat dicari dengan cara sbb:

$$t_n y(x_n) = 1 \qquad (2.4)$$

$$t_n \left(\sum_{m \epsilon S} a_m^* t_m k\,(x_n, x_m) + b \right) = 1 \;\rightarrow b$$

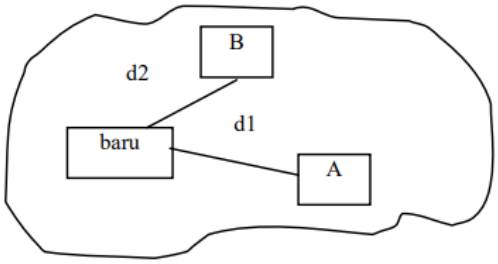
$$= \frac{1}{N} \sum_{m \epsilon S} (t_n - \sum_{m \epsilon S} a_m^* t_m k\,(x_n, x_m)) \qquad (2.5)$$

dimana S adalah himpunana indeks dari support vectors, dan N_s adalah jumlah semua support vectors

1.6.4. K-Nearest Neighbor

Algortima K-Nearest Neighbor (KNN) adalah merupakan sebuah metode untuk melakukan klasifikasi terhadap objek baru berdasarkan (K) tetangga terdekatnya. KNN termasuk algoritma supervised learning, dimana hasil dari query instance yang baru, diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Kelas yang paling banyak muncul yang akan menjadi kelas hasil klasifikasi (Kusrini dkk., 2009). K-Nearest Neighbor adalah suatu pendekatan untuk menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. Illustrasi kedekatan kasus pada Gambar 2.3. memberikan gambaran tentang proses mencari solusi terhadap seorang pasien baru dengan menggunakan mengacu pada solusi dari pasien terdahulu. Untuk mencari kasus pasien mana yang akan digunakan, maka dihitung kedekatan antara kasus pasien baru dengan semua kasus pasien lama. *K-Nearest Neighbor* (kNN) adalah suatu metode yang menggunakan algoritma *supervised* dimana hasil dari *query instance* yang baru diklasifikan berdasarkan mayoritas dari kategori pada kNN. Tujuan dari algoritma kNN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan training samples. Dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada kNN (Fansyuri, 2020).

Gambar 3. Ilustrasi Kedekatan Kasus



Langkah-langkah pada algoritma KNN:

1. Tentukan jumlah tetangga (K) yang akan digunakan untuk pertimbangan penentuan kelas.
2. Hitung jarak dari data baru ke masing-masing data point di dataset.
3. Ambil sejumlah K data dengan jarak terdekat, kemudian tentukan kelas dari data baru tersebut.

Untuk mencari dekat atau jauhnya jarak antar titik pada kelas k dihitung menggunakan jarak Euclidean. Jarak Euclidean atau *Euclidean*

Distance adalah formula untuk mencari jarak antara 2 titik dalam ruang dua dimensi. Berikut rumus untuk menghitung jarak Euclidean:

$$d\ (x,y) = \sqrt{\sum_{i=1}^{n} \left(x_{training}^{i} - y_{testing}^{i}\right)^2}$$

Keterangan :

d (x,y)	= Jarak
$x_{training}^{i}$	= Data Training
$y_{testing}^{i}$	= Data Testing
i	= Variabel Data
n	= Dimensi Data

1.6.5. Kinerja Klasifikasi

Evaluasi akurasi model klasifikasi menggunakan metode *confusion matrix* untuk mengetahui apakah model yang digunakan baik atau tidak. Berikut ilustrasi *confusion matrix* menurut (C. Ortega, 2020).

Tabel 1. Ilustrasi confusion matrix

Kelas	Aktual Positif (1)	Aktual Negatif (0)
Prediksi Positif (1)	TP (Benar)	FP (Salah)
Prediksi Negatif (0)	FN (Salah)	TN (Benar)

Dimana:
TP: Model memprediksi benar dan itu benar
TN: Model memprediksi negatif dan itu benar
FP: Model memprediksi positif dan itu salah
FN: Model memprediksi negatif dan itu salah

Berdasarkan nilai (TP), (TN), (FP), dan (FN), diperoleh nilai *Accurasi, Precision*, *Recall,* dan *F-Measure* (Natakusumah & Ernastuti, 2022).

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (2.7)$$

$$Precision = \frac{(TP)}{(TP + FP)} \quad (2.8)$$

$$Recall = \frac{(TP)}{(TP + FN)} \quad (2.9)$$

$$F - Measure = 2\frac{(Recall \times Precision)}{(Recall + Precision)} \quad (2.10)$$

Nilai akurasi menggambarkan seberapa akurat sistem dapat mengklasifikasikan data secara benar dengan keseluruhan data. Nilai presisi

menggambarkan jumlah data kategori positif yang diklasifikasikan secara benar dibagi dengan total data yang diklasifikasi positif. *Recall* menunjukkan berapa persen data kategori positif yang terklasifikasi dengan benar oleh sistem.

Nilai AUC memiliki rentang antara 0.5 sampai dengan 1. Interpretasi nilai AUC dapat diklasifikasikan menjadi lima bagian yang berbeda yaitu Akurasi salah, tingkat akurasi lemah, tingkat akurasi sedang, tingkat akurasi tinggi, dan tingkat akurasi tinggi. Rumus AUC sebagai berikut:

$$AUC = \frac{\sum_{i=1}^{n^+}\sum_{j=1}^{n^-} 1_{f(x_i^+)>f(x_j^-)}}{n^+n^-} \quad (2.11)$$

Dimana:

$f(.)$: Nilai suatu fungsi
x^+ : sampel positif
x^- : sampel negatif
n^+ : jumlah sampel positif
n^- : jumlah sampel negative

1.6.6. Covid ‘19

Penyakit Coronavirus tahun 2019, dikenal sebagai COVID-19, adalah penyakit menyebar cepat yang disebabkan oleh Sindrom Pernafasan Akut Parah Coronavirus 2 (SARS-CoV2). COVID-19 sekarang dianggap pandemi yang telah mempengaruhi negara-negara di semua yang dihuni benua. Sejak kasus pertama COVID- 19 dilaporkan di Wuhan, China, pada Desember 2019, jumlah kematian di seluruh dunia telah meningkat dengan cepat.

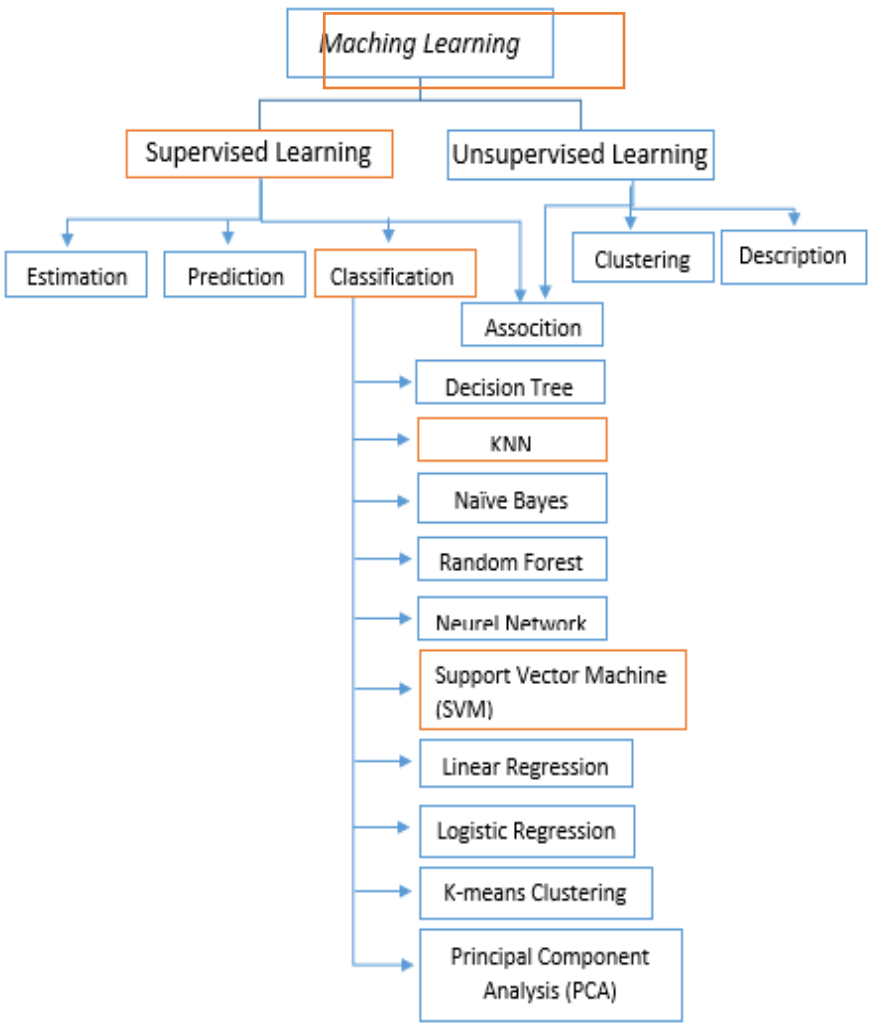
Karena itu infeksi tinggi dan angka kematian, pemerintah miliki mengimplementasikan berbagai kebijakan yang ditujukan untuk mengurangi penyebaran virus ini dan dampaknya. Seperti itu tindakan dimulai dengan perintah pemerintah China untuk Karantina Wuhan pada 23 Januari 2020, untuk sebagian besar Baru-baru ini, beberapa negara menyatakan keadaan darurat dan menerapkan karantina yang ketat dan menjaga jarak social (Bullock dkk., 2020). Sebuah coronavirus baru telah diidentifikasi dengan kasus pasien pertama yang dikonfirmasi pada bulan Desember 2019 di kota Wuhan, provinsi Hubei, di Cina. Sejak itu, jumlah kasus yang dikonfirmasi telahmeningkat drastis (Alimadadi dkk., 2020).

Coronavirus merupakan suatu kelompok virus yang dapat menyebabkan penyakit pada hewan atau manusia. Beberapa jenis corona virus diketahui menyebabkan infeksi saluran nafas pada manusia mulai dari batuk pilek hingga yang lebih serius seperti *Middle East Respiratory Syndrome* (MERS) dan *Severe Acute Respiratory Syndrome* (SARS).

Coronavirus jenis baru yang ditemukan menyebabkan penyakit COVID-19, Orang dapat tertular COVID-19 dari orang lain yang terjangkit virus ini. COVID-19 dapat menyebar dari orang ke orang melalui percikan-percikan dari hidung atau mulut yang keluar saat orang yang terjangkit COVID-19 batuk atau mengeluarkan napas. Percikan-percikan ini kemudian jatuh ke benda- benda dan permukaan-permukaan disekitar. Orang yang menyentuh benda atau permukaan tersebut lalu menyentuh mata, hidung atau mulutnya, dapat terjangkit COVID-19. Penularan COVID-19 juga dapat terjadi jika orang menghirup percikan yang keluar dari batuk atau napas orang yang terjangkit COVID-19 dan COVID-19, penyakit yang disebabkan oleh virus SARS-CoV-2, telah dinyatakan pandemi oleh Organisasi Kesehatan Dunia (WHO)(Li dkk., 2020).

1.6.7. Kerangka Konseptual

Gambar 4. Kerangka Konseptual



BAB II
METODOLOGI PENELITIAN

2.1 Data dan Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder, yaitu berupa data pasien COVID-19 tahun 2020-2021 yang diperoleh dari Dinas Kesehatan Kota Makassar. Data tersebut merupakan data tentang faktor-faktor yang mempengaruhi ketahanan hidup pasien Covid-19 di Kota Makassar.

Data Covid 19 dibagi menjadi data *training* dan *testing*.

2.2 Varibel Penelitian

Variabel yang akan digunakan pada penelitian ini adalah :

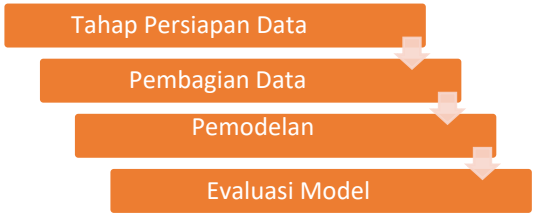
Tabel 2. Variabel

Variabel Penelitian	Kode	Nama Variabel	Skala Pengukuran	
Variabel Respon	y	Status	1	Mati
			0	Hidup
Variabel Prediktor	x_1	Jenis kelamin	1	Perempuan
			2	Laki-kaki
	x_2	Usia	Tahun	
	x_3	Komorbid	1	Positif dengan komorbid
			2	Positif dengan non komorbid
	x_4	Lama Perawatan	Hari	

2.3 Tahapan Penelitian

Penelitian ini berfokus pada pengembangan model *data mining* dengan membandingkan klasifikasi *Support Vector Machine* dan *K-Nearest Neigbor.* Adapun metode analisis disusun sebagai berikut:

Gambar 5. Tahapan Penelitian



1. Tahap Persiapan Data :
    1. Identifikasi data yang terkumpul dalam basis data yang tersedia di Dinas Kesehatan Kota Makassar
    2. Melakukan seleksi atau pemilihan data yang relevan dengan analisis yang dilakukan
    3. *Preparation* data dan transformasi data yang terdiri dari :
        a. Pembersihan data diterapkan untuk menghilangkan *noise* dan memperbaiki data yang tidak konsisten
        b. Reduksi data digunakan untuk mengurangi ukuran data. Misalnya menggabungkan, menghilangkan fitur yang berlebihan atau mengelompokkan transformasi.
    4. Transformasi data yang telah dipilih menjadi *coding* agar data tersebut sesuai dengan proses *data mining*

2. Tahap Pembagian Data

    Membagi data ke dalam data latih dan data uji. Pada penelitian ini pembagian data menggunakan sampel acak sederhana

3. Tahap Pemodelan

    Pada bagian ini dijelaskan tentang langkah-langkah eksperimen meliputi cara pemilihan arsitektur yang tepat dari model atau metode yang diusulkan sehingga didapatkan hasil yang dapat membuktikan bahwa metode yang digunakan adalah tepat. Menggunakan algoritma *Support Vector Machine* dan *K-Nearest Neighbor*

4. Evaluasi Model

    Pada bagian ini dijelaskan tentang evaluasi dan validasil hasil penerapan metode pada penelitian yang dilakukan predeksi dengan data uji dengan model klasifikasi yang diperoleh dari tahap pemodelan *Support Vector Machine* dan *K-Nearest Neighbor.*