BAB II
KAJIAN LITERATUR

2.1 Tinjauan Pustaka

Pada subbab ini, akan dijelaskan tinjauan pustaka yang berkaitan dengan penelitan yang akan dilakukan. Suatu penelitian harus memiliki dukungan melalui hasil-hasil penelitian yang telah ada sebelumnya. Terutama penelitian sebelumnya yang berkaitan dengan penelitian yang dilakukan saat ini mencakup prediksi dengan data mining untuk mengetahui hasil ujian siswa dan faktor-faktor yang mempengaruhinya. Beberapa penelitian sebelumnya (Penulis urutkan dari tahun rendah ke tahun tinggi) yang membahas hal yang sama antara lain.

(Uricar *et al.*, 2016) melakukan penelitian dengan judul "Structured Output *SVM* Prediction of Apparent Age, Gender and Smile From Deep *Feature*s" . Penelitian ini mengusulkan *SVM* keluaran terstruktur untuk memprediksi aplikasi
usia orang tua serta jenis kelamin dan senyum dari gambar wajah tunggal yang diwakili oleh Deep *Feature*s. Penelitian ini mengajukan masalah perkiraan usia orang tua sebagai sebuah instance dari pengklasifikasi *SVM* keluaran terstruktur multi-kelas yang diikuti oleh perbaikan nilai yang diharapkan softmax. Prediksi jenis kelamin dan senyum diperlakukan sebagai masalah klasifikasi biner. Solusi yang diusulkan pertama-tama mendeteksi wajah dalam gambar dan kemudian mengekstrak fitur dalam dari gambar yang dipotong di sekitar wajah yang terdeteksi. Penelitian menggunakan jaringan neural konvolusional dengan arsitektur VGG-16 untuk mempelajari Deep *Feature*s. Jaringan tersebut telah dilatih sebelumnya pada database ImageNet dan kemudian disetel pada *Dataset* IMDB-WIKI dan ChaLearn 2015 LAP. Kami memvalidasi metode kami pada *Dataset* ChaLearn 2016 LAP . *SVM* keluaran terstruktur penelitian ini dilatih hanya pada data LAP ChaLearn 2016. Penelitian ini mencapai hasil yang sangat baik untuk prediksi usia dan jenis kelamin serta senyuman.

(Bydžovská, 2016) melakukan penelitian dengan judul "A Comparative Analysis of Techniques for Predicting Student Performance" penelitian ini bertujuan untuk memprediksi nilai akhir siswa dalam mata pelajaran tertentu dengan menggunakan teknik data mining. Menggunakan dua pendekatan yang divalidasi pada 138 mata kuliah yang ditawarkan kepada mahasiswa Fakultas Informatika Universitas Masaryk antara tahun 2010 dan 2013. Pendekatan pertama didasarkan pada algoritma klasifikasi dan regresi yang mencari pola dalam data terkait studi dan juga data. tentang perilaku sosial siswa. Penelitian ini membuktikan bahwa karakteristik perilaku sosial siswa meningkatkan prediksi untuk seperempat mata pelajaran. Pendekatan kedua didasarkan pada teknik penyaringan kolaboratif. Kami memprediksi nilai akhir berdasarkan prestasi sebelumnya dari siswa serupa. Hasilnya menunjukkan bahwa kedua pendekatan tersebut mencapai hasil rata-rata yang serupa dan dapat digunakan secara bermanfaat untuk prediksi nilai akhir siswa. Pendekatan pertama mencapai hasil yang jauh lebih baik untuk kursus dengan jumlah siswa yang sedikit. Sebaliknya, pendekatan kedua mencapai hasil yang lebih baik secara signifikan untuk kursus matematika. Penelitian ini menggunakan beberapa metode data mining. Penelitian ini menyimpulkan *SVM* adalah metode terbaik dalam penelitian ini.

Kuzilek, et al, (2017) melakukan penelitian dengan judul "Data Descriptor: Open University Learning Analytics *Dataset*". Penelitian ini bertujuan untuk membahas mengenai *Dataset Oulad* yang akan penulis gunakan untuk melakukan penelitian ini. Penelitian ini menjelaskan apakah itu *Dataset Oulad* , bagaimana cara mendapatkannya, apa saja isi *Dataset* tersebut. Dalam penelitian ini juga dijelaskan bagaimana *Dataset* tersebut dapat dikumpulkan sambil tetap menjaga privasi dari pemberi data. Kesimpulannya adalah *Dataset Oulad* merupakan *Dataset* yang sangat bagus untuk digunakan untuk penelitian dan sudah tersertifikasi Open Data Institute. Dalam hal ini penulis melihat pemecahan untuk masalah kurangnya jumlah data yang dilakukan pada penelitian sebelumnya dan sangat tertarik dengan keunikan *Dataset* ini yang berisikan data pembelajaran online. Kebetulan pada masa pademi corona ini pembelajaran online sedang melonjak popularitasnya.

Penelitian yang dilakukan oleh GritNet: Student Performance Prediction with Deep Learning (Kim, Vizitei and Ganapathi, 2018). Penelitian ini memprediksi kinerja siswa di masa depan saat mereka berinteraksi dengan kursus online. Prediksi tahap awal dari kinerja siswa di masa depan bisa menjadi sangat penting untuk memfasilitasi intervensi pendidikan yang tepat waktu selama kursus. Dalam penelitian ini, menggunakan algoritma berbasis deep learning baru, disebut GritNet, yang dibangun di atas memori jangka pendek dua arah (BLSTM). Hasil menunjukkan GritNet tidak hanya secara konsisten mengungguli metode berbasis regresi logistik standar, tetapi peningkatan tersebut secara substansial diucapkan dalam beberapa minggu pertama ketika prediksi yang akurat paling menantang. Yang menarik adalah GritNet itu dibagun dari JST yang merupakan dasar dari *Neural Tangent Kernel* yang akan digunakan dalam penelitian ini.

Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika (Putri and Waspada, 2018). Penelitian ini adalah memanfaatkan data tentang mahasiswa yang lulus dengan mengolahnya menggunakan data mining untuk mendapatkan informasi berupa prediksi kelulusan mahasiswa. Metode yang akan digunakan adalah metode pohon keputusan yang dibangun dengan algoritma C4.5 disertai dengan algoritma error-based pruning untuk proses pemotongan pohon keputusan. Kriteria yang akan digunakan adalah jenis kelamin, asal daerah, IPK, dan TOEFL. Dalam penerapannya, algoritma C4.5 dapat digunakan untuk menghasilkan prediksi kelulusan dengan nilai rata-rata precision 63.93%, *Recall* 60.73%, dan akurasi 60.52%. Setelah pohon keputusan dipotong dengan menggunakan metode error-based pruning, didapatkan hasil yang lebih baik. Pohon yang dipotong dengan menggunakan nilai confidence 0,4 menghasilkan precision 70.70%, *Recall* 50.65%, dan akurasi 61.57%. Sedangkan pohon yang dipotong dengan menggunakan menggunakan nilai confidence 0,25 menghasilkan precision 73.77%, *Recall* 48.84%, dan akurasi 62.44%. Penelitian ini menunjukkan dengan data learning yang sedikit maka hasil precision, *Recall* dan akurasi tidak begitu memuaskan.

Jacot, et al, (2018) melakukan penelitian dengan judul "*Neural Tangent Kernel: Convergence and generalization in neural networks*". Tujuan dari penelitian tersebut adalah memperkenalkan alat baru untuk mempelajari *JST, Neural Tangent Kernel (NTK)*, yang menjelaskan *dinamika lokal JST* selama *penurunan gradien.* Ini mengarah ke koneksi baru antara *ANN pelatihan* dan *metode kernel.* Hasil dari penelitian adalah kinerja *JST* dapat secara drastis meningkat saat lebar lapisannya bertambah besar. *NTK* menjelaskan evolusi jaringan saraf di bawah penurunan gradien dalam ruang fungsi. Perspektif ini adalah pemahaman tentang bagaimana jaringan saraf berkembang dalam ruang parameter, karena *NTK* didefinisikan dalam istilah gradien keluaran *JST* sehubungan dengan *parameternya.* Dalam batas lebar yang tak terhingga, hubungan antara kedua perspektif ini menjadi sangat menarik. Peneliti ini dan rekannya juga menjelaskan dalam penelitiannya dan menerangkan di video youtube bahwa *NTK* sangat bagus dikombinasikan dengan *SVM*.

(Yaacob *et al.*, 2019) melakukan penelitian dengan judul *"Supervised data mining approach for predicting student performance"*. Penelitian ini bertujuan untuk mengembangkan model prediktif menggunakan algoritma klasifikasi memprediksi kinerja siswa menjadi siswa yang sangat baik atau tidak unggul tergantung pada hasil kinerja akademis mereka melalui data mining. Empat pengklasifikasi seperti *Decision Trees, Naïve Bayes, KNearest Neighbor, Logistic Regression* diadopsi dalam memprediksi kinerja siswa dan dikategorikan mereka sangat baik atau tidak sangat baik. Dalam studi ini langkah-langkah yang terlibat dalam metodologi untuk mengembangkan *model prediktif* menggunakan data mining diimplementasikan mengikuti *model CRISP-DM (Proses Standar Lintas Industri untuk Data Mining)* . Proses *CRISP-DM* merupakan pendekatan siklik yang terdiri dari enam langkah. Langkah pertama adalah memahami aktivitas bisnis dan masalah di mana proses tersebut melibatkan transformasi bisnis masalah memprediksi kinerja siswa menjadi masalah *data mining.* Kemudian, langkah kedua melibatkan data analisis termasuk pengumpulan dan pengenalan data mentah. Selanjutnya, persiapan data. Langkah keempat adalah data pemodelan yang melibatkan beberapa *algoritma prediktif*

dikembangkan termasuk *K-NN, Naïve Bayes, Decission Tree dan Model Regresi Logistik.* Setelah model dikembangkan, langkah terakhir selanjutnya adalah model evaluasi dan penyebaran. Hasil dari penelitian ini adalah Hasilnya menunjukkan bahwa pengklasifikasi *Naïve Bayes* mengungguli algoritma lain dibandingkan *Decision Tree, k-NN, dan Logistic Regression* dengan pengklasifikasi yang akurat dan komprehensif. Penelitian ini membuktikan bahwa prediksi kinerja siswa itu penting dilakukan untuk universitas untuk meningkatkan kinerja pengajaran mereka.

(Irmawati, Zainuddin and Yuyun, 2020) melakukan penelitian dengan judul "*Data Mining Untuk Penentuan Model Tingkat Kesuksesan Kelulusan Murid SMA Pada Perguruan Tinggi Negeri: Studi Kasus Di Iain Bone*". Penelitian ini bertujuan untuk memprediksi tingkat keberhasilan kelulusan murid pada Institut Agama Islam Negeri (IAIN) Bone dengan menggunakan metode algoritma Naive Bayes dan C.45. Algoritma yang diusulkan untuk mempredikasi kriteria apa saja yang menjadi penentu kelulusan murid pada penerimaan mahasiswa baru jalur mandiri tahun 2018. Penelitian ini nantinya akan memberikan informasi kepada IAIN Bone mengenai klasisifkasi tingkat kesuksesan kelulusan murid SMA yang diterima dan sekaligus memberikan informasi kepada murid SMA yang akan lulus dari sekolah mengenai fakto-faktor apa saja yang menjadi penentu tingkat kesuksesan kelulusan pada IAIN Bone. Tujuh kriteria yang digunakan sebagai variabel pendukung dalam melakukan analisis. Kriteria tersebuat adalah Tahun Lulus, Pendidikan Orang Tua, Penghasilan Orang Tua, Nilai Ujian Nasional, Nilai Tes, Nilai Wawancara dan Nilai Baca Tulis Huruf Qur'an (BTHQ). *Dataset* dalam penelitian ini bersumber dari Database Sistem Informasi Akademik (SISFO) IAIN Bone dari Tiga sekolah yaitu SMA 4 Watampone, MAN 1 Bone dan SMKN 1 Watampone. Hasil pengujian menunjukkan bahwa Nilai BTHQ menjadi syarat utama kelulusan murid SMA jalur mandiri pada IAIN Bone sesuai dengan hasil olahan data training sebanyak 170 dan olahan data testing sebanyak 10. Kedua algoritma menghasilkan Nilai Precision sebesar 100%, *Recall* 100%, Accuracy 100%. Selain itu, ditemukan pola sequential baru dengan melakukan pengujian ulang berdasarkan hasil urutan kejadian dari variabel Nilai Ujian Nasional, Nilai Tes, Nilai

Wawancara dan Nilai Baca Tulis Huruf Qur'an (BTHQ). Penelitian ini memperoleh hasil akurasi yang tinggi. Ini yang menjadai motivasi penulis untuk dapat menghasilkan nilai precision yang tinggi.

Tomasevic, et al, (2020) melakukan penelitian dengan judul "*An overview and comparison of supervised data mining techniques for student exam performance prediction"*. Tujuan penelitian tersebut yaitu untuk menyediakan analisa perbandingan dari teknik *supervised machine learning* yang diaplikasikan untuk menyelesaikan prediksi perfoma ujian murid. tugas prediksi kinerja ujian siswa, yaitu menemukan siswa pada "risiko tinggi" untuk *Drop Out*, dan memprediksi pencapaian masa depan mereka, seperti misalnya, hasil ujian akhir. Data uji menggunakan Kumpulan data *Open University Learning Analytics (OULAD)* berisi *subset data siswa Open University (OU)* dari tahun 2013 dan 2014. Data tersebut mencakup data demografi siswa dan data interaksi dengan *Virtual Learning Environment (VLE)* universitas. *Dataset* berisi 22 modul-presentasi dengan 32.593 siswa terdaftar dan tersedia secara gratis di https://analyse.kmi.open.ac.uk/open_*Dataset*. OULAD telah disertifikasi oleh Open Data Institute (http://theodi.org/). Tahapan dari penelitian tersebut adalah menyiapkan *Dataset* lalu melakukan proses *pre-processing*. Melakukan pengolahan data dan mendapatkan perbandingan dari 8 algoritma dimana *SVM* termasuk salah satu di dalamnya. Hasil dari penelitian tersebut adalah Untuk tugas klasifikasi dan regresi, Hasil ketepatan prediksi tertinggi secara keseluruhan diperoleh dengan metode jaringan saraf tiruan dengan memasukkan data keterlibatan siswa dan data performa masa lalu siswa, sedangkan penggunaan data demografis tidak menunjukkan pengaruh yang signifikan pada ketepatan prediksi. Untuk memanfaatkan potensi penuh dari prediksi kinerja ujian siswa, Disimpulkan bahwa fungsionalitas akuisisi data memadai dan interaksi siswa dengan lingkungan belajar merupakan prasyarat untuk memastikan jumlah data yang cukup untuk analisis. Pada penelitian ini tingkat akurasi dari *SVM* masih dibawah jaringan saraf tiruan.

(Wakelam et al., 2020) melakukan penelitian dengan judul "*The potential for student performance prediction in small cohorts with minimal available attributes*" .

Penelitian ini bertujuan untuk mencari jawaban apakah mungkin dan berguna untuk memprediksi kinerja siswa pada kursus yang terdiri dari yang relatif kecil kelompok siswa, di mana sekumpulan data siswa yang sangat terbatas tersedia untuk analisis dan seberapa berguna analisis ini untuk memberikan kesempatan kepada para pengelola kursus membuat intervensi pendukung tepat waktu pada titik-titik yang sesuai selama modul. Hasil dari penelitian ini adalah Akurasi prediksi rata-rata di semua algoritma yang digunakan adalah 67%, dengan akurasi prediksi KNN dan RF antara 66% dan 75%. Di penelitian ini penulis melihat jumlah *Dataset* yang sedikit justru menghasilkan tingkat akurasi prediksi yang rendah dibandingkan penelitian yang menggunakan *Dataset* besar seperti OULAD yang dilakukan oleh (Tomasevic, Gvozdenovic and Vranes, 2020).

Pada bagian tinjauan pustaka ini juga dijabarkan mengenai teori-teori yang digunakan dalam penelitian ini dalam beberapa sub bab sebagai berikut.

2.1.1 Data Mining

Data Mining dijabarkan sebagai proses untuk menemukan hubungan, pola dan tren baru yang bermakna dengan menyaring data yang sangat besar, yang tersimpan dalam penyimpanan, menggunakan teknik pengenalan pola seperti teknik Statistik dan Matematika(Kamagi & Hansun, 2014). *Data mining* adalah sebuah proses percarian secara otomatis informasi yang berguna dalam tempat penyimpanan data berukuran besar. Ada beberapa Istilah lain dalam *Data Mining*. Istilah lain yang sering digunakan diantaranya *knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, dan business intelligence.* (Windarto, 2017).

Connolly dan Begg (2010) menyatakan bahwa *data mining* adalah suatu proses ekstraksi atau penggalian data yang belum diketahui sebelumnya, namun dapat dipahami dan berguna dari *database* yang besar serta digunakan untuk membuat suatu keputusan bisnis yang sangat penting.

2.1.2 Pengolahan Data Mining

PengolahanData mining terdiri dari beberapa metode pengolahan, yaitu (Kamagi and Hansun, 2014)

a. *Predictive modelling* merupakan pengolahan data mining dengan melakukan prediksi . Contoh algoritmanya *Linear Regression, Neural Network, Support Vector Machine*, dan lain- lain.
b. *Association (Asosiasi)* merupakan teknik yang mempelajari hubungan antar data. Contoh penggunaannya seperti untuk menganalisis perilaku mahasiswa yang datang terlambat. Contoh *algoritmanya FP-Growth, A Priori,* dan lain-lain.
c. *Clustering (Klastering)* atau pengelompokkan merupakan teknik untuk mengelompokkan data ke dalam suatu kelompok tertentu. Contoh algoritmanya *K-Means, K-Medoids, Self-Organitation Map (SOM), Fuzzy CMeans*, dan lain-lain.
d. *Classification* merupakan teknik mengklasifikasikan data. Perbedaannya dengan *metode clustering* terletak pada data. Contoh *algoritma ID3* dan *K Nearest Neighbors.*

Adapun Karakteristik *data mining* menurut (Lorena et al., 2014) sebagai berikut:

a. Mempunyai hubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
b. Menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih dipercaya.
c. Membuat keputusan yang kritis, terutama dalam strategi.

2.1.3 Konsep Prediksi

Prediksi adalah proses untuk memperkirakan secara sistematis tentang sesuatu yang paling mungkin terjadi di masa depan berdasarkan data masa lalu dan sekarang yang dimiliki, agar persentase kesalahannya (selisih antara sesuatu yang terjadi dengan

hasil perkiraan) dapat diperkecil. Dalam data mining, konsep prediksi merupakan *pemodelan* yang dilakukan menggunakan data sampel yang diketahui nilai aributnya untuk memperkirakan nilai atribut dari target tertentu. (Werdiningsih, et al., 2020)

Berdasarkan caranya, prediksi bisa berdasarkan metode ilmiah ataupun subjektif belaka. Ambil contoh, prediksi cuaca hujan atau tidak selalu berdasarkan data dan informasi terbaru yang didasarkan pengamatan termasuk oleh satelit. Namun, prediksi seperti pertandingan hasil pertandingan sepakbola, umumnya berdasarkan pandangan subjektif dengan sudut pandang sendiri yang memprediksinya. (Binti Hamzah *et al.*, 2020)

Menurut penelitian (Tomasevic, Gvozdenovic and Vranes, 2020) sekarang ini sudah banyak metode data mining yang dapat digunakan untuk memprediksi dengan tingkat keakuratan yang tinggi dimana salah satunya *SVM* yang akan digunakan penulis dalam penelitian ini.

2.1.4 Pendekatan Dalam Data Mining

Data Mining merupakan kumpulan dari kegiatan yang meliputi pengumpulan dan pemakaian data masa lalu untuk menemukan pola atau hubungan dalam data yang berukuran besar. Output pada data mining tersebut dapat dijadikan pengambilan keputusan dimasa depan. Metode ini merupakan gabungan dari 4 disiplin ilmu yaitu *visualisasi, statistik, basis data* dan *machine learning.* (Sodik, Dwi and Kharisudin, 2020).

Menurut (Cattral, Oppacher and Deugo, 2001) dalam *data mining* dikenal *Supervised learning dan unsupervised learning,* dimana akan dijabarkan seperti berikut ini :

1. *Supervised learning*

   Supervised learning adalah tipe Machine Learning dimana model ini menyediakan *training* data berlabel. Dalam bahasa Indonesia, arti *Supervised learning* adalah pembelajaran mesin yang diawasi karena memiliki "label" yang menunjukan mana bagian "hasil". Label digunakan untuk kolom

jawaban. Contohnya label (secara mudah kita melihat sebagai nama kolom jawaban), misal: “result”, “jawaban”, atau “hasil”. Dalam Supervised learning terdapat data training dan data testing.

2. Unsupervised Learning

   Unsupervised learning dalam bahasa Indonesia adalah “pembelajaran tanpa pengawasan”. *Unsupervised* bertujuan untuk mengidentifikasi pola yang memiliki makna dalam data. Jika *Supervised Learning* belajar dari data dengan label, maka di *Unsupervised* mesin harus belajar dari kumpulan data tanpa label. Karena tidak adanya petunjuk, salah satu cara untuk mendapatkan memprediksinya adalah dengan menggunakan *cluster* (salah satu algoritma dalam *Unsupervised Learning*)(Dickyibrohim, 2019).

Dalam *supervised learning*, tujuannya adalah untuk mempelajari pemetaan dari input ke output yang nilai benarnya diberikan oleh seorang supervisor. Dalam unsupervised learning, tidak ada pembimbing dan hanya memiliki data masukan. Tujuannya adalah untuk menemukan keteraturan dalam masukan. Ada struktur pada ruang masukan sedemikian rupa sehingga pola tertentu lebih sering muncul daripada yang lain, dan ingin melihat apa yang umumnya terjadi dan apa yang tidak. Dalam statistik, ini disebut *estimasi kepadatan.* (Alpaydin, 2014)

2.1.5 Algoritma Supervised Learning

Beberapa contoh macam *Supervised learning* yang tersedia menurut (Han, Kamber and Pei, 2012) adalah :

1. Decision Tree: Sebuah *Decision Tree* adalah alat pendukung keputusan yang menggunakan *grafik treelike* atau model keputusan dan konsekuensi yang mungkin mereka, termasuk hasil chanceevent, biaya sumber daya, dan utilitas
2. Naïve Bayes: *Naïve Bayes* adalah keluarga dari pengklasifikasi probabilistik sederhana berdasarkan menerapkan teorema Bayes 'dengan (naif) asumsi independensi yang kuat antara fitur.

3. Ordinary Least Squares Regression Ordinary Least Squares Regression Linear mengacu pada jenis model yang digunakan agar sesuai dengan data, sementara kuadrat mengacu pada jenis *kesalahan metrik* Anda meminimalkan lebih.
4. Support Vector Machines: *SVM* adalah algoritma klasifikasi biner. Mengingat satu set poin dari 2 jenis di N tempat dimensi, *SVM* menghasilkan (N - 1) hyperlane dimensi untuk memisahkan titik-titik menjadi 2 kelompok

2.1.6. Algoritma Unsupervised Learning

Beberapa contoh macam *Unsupervised learning* yang tersedia menurut (Bert, 2019) adalah :

1. Pengelompokan K-means. Dalam daftar algoritma pembelajaran tanpa pengawasan, ini mungkin metode yang paling sederhana. Seperti namanya, idenya adalah untuk mendefinisikan cluster berdasarkan pusat K. Pusat K ditempatkan sedemikian rupa untuk memaksimalkan perbedaan atau jarak antara masing-masing, dan data apa pun ditempatkan ke grup dengan pusat-K terdekat.
2. Principal Component Analysis (PCA). PCA adalah prosedur dimana data diklasifikasikan menjadi satu set komponen yang tidak terkait yang disebut komponen utama. Ini dilakukan dengan bantuan transformasi ortogonal, yang merupakan transformasi linier dari ruang vektor yang mempertahankan panjang vektor
3. Aturan Asosiasi. Ini adalah teknik tanpa pengawasan lain yang umum digunakan untuk mengetahui hubungan antara komponen dalam database yang luas. Ini adalah metode berbasis aturan. Misalnya, Anda dapat menemukan hubungan menarik antara produk yang dibeli dan penggunaan mesin PoS dalam data penjualan supermarket.

2.1.7 Tahapan Dalam Data Mining

Untuk mendapatkan hasil dari data mining yang memuaskan kita harus mengikuti tahapan yang sudah dilakukan peneliti yang telah berhasil sebelumnya.

Menurut (Eska, 2016), tahapan dalam d*ata mining* diurutkan menjadi :

Tahapan *data mining* dibagi menjadi enam bagian yaitu :

- Pembersihan data (data cleaning),

  sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus. Proses cleaning mencakup antara lain membuang 4 *duplikasi data,* memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan *proses enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau *informasi eksternal*.

- Integrasi data *(data integration)*

  Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru. Tidak jarang data yang diperlukan untuk data mining tidak hanya berasal dari satu database tetapi juga berasal dari beberapa database atau file teks. Integrasi data dilakukan pada atribut-aribut yang mengidentifikasikan *entitas-entitas* yang unik seperti atribut nama, jenis produk, nomor pelanggandan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.

- Seleksi Data (*Data Selection*)

  Data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database. Sebagai contoh, sebuah kasus yang meneliti faktor kecenderungan orang

membeli dalam kasus *market basket analysis*, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

- Transformasi data *(Data Transformation)*

  Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining. Beberapa metode *data mining* membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan clustering hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa *interval.* Proses Ini sering disebut *transformasi data.*

- *Proses mining.*

  adalah sebuah proses yang paling utama pada saat metode diterapkan untuk mencari pengetahuan tersembunyi dan berharga dari data.

- Evaluasi pola *(pattern evaluation)*,

  Untuk mengidentifikasi pola-pola menarik kedalam knowledge based yang ditemukan. Dalam tahap ini hasil dari teknik data mining berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai.

- Presentasi pengetahuan *(knowledge presentation)*,

  Merupakan penyajian dan visualisasi pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahap terakhir dari proses data mining adalah bagaimana memformulasikan keputusan atau 5 aksi dari hasil analisis yang didapat. Ada kalanya hal ini harus melibatkan orang-orang yang tidak memahami data mining. Karenanya presentasi hasil data mining dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan dalam proses data mining. Dalam presentasi ini,visualisasi juga bisa membantu mengkomunikasikan hasil data mining.

2.1.8 *SVM*

SVM merupakan salah satu metode terbaik yang bisa dipakai dalam permasalahan klasifikasi dan prediksi. Konsep *SVM* bermula dari masalah klasifikasi dua kelas sehingga membutuhkan *training set* positif dan negatif. *SVM* berusaha menemukan *hyperplane* (pemisah) terbaik untuk memisahkan ke dalam dua kelas dan memaksimalkan *margin* antara dua kelas tersebut. Pada beberapa kasus, data tidak bisa diklasifikasi menggunakan *metode linier SVM,* sehingga dikembangkan fungsi kernel untuk mengklasifikasikan data dalam bentuk *nonlinier.* (Arif Pratama et al, 2018)

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di Annual Workshop on Computational Learning Theory. Konsep dasar *SVM* sebenarnya merupakan kombinasi harmonis dari teoriteori komputasi yang telah ada puluhan tahun sebelumnya, seperti *margin hyperplane* (Duda & Hart tahun 1973, Cover tahun 1965, Vapnik 1964, dsb.), *kernel* diperkenalkan oleh Aronszajn tahun 1950, dan demikian juga dengan konsep-konsep pendukung yang lain. Akan tetapi hingga tahun 1992, belum pernah ada upaya merangkaikan komponen-komponen tersebut (Nugroho, 2007).

Persamaan *SVM* dasar adalah :

$$\int(x) = w^t \emptyset(x) + b \qquad (2.1)$$

Dimana :

b = Bias

x = $(x_1, x_2, \ldots . x_D)^T$ = Variabel Input

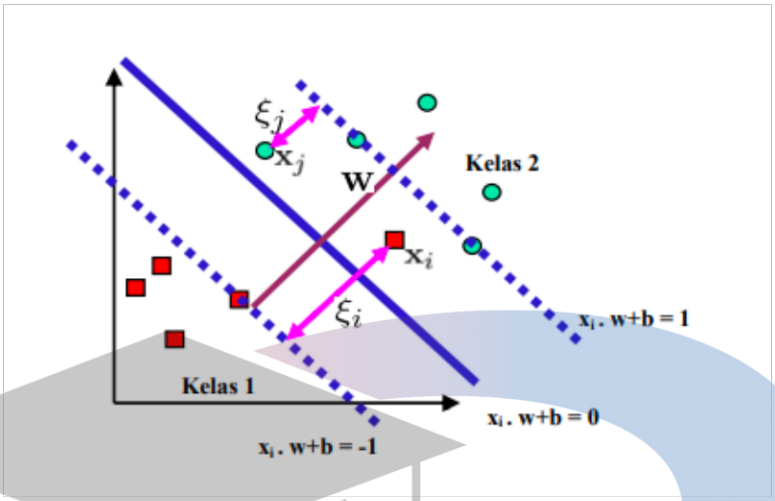
w = $(w_0, w_1, \ldots . w_D)^T$ = Parameter Bobot

$\emptyset\ (x)$ = Fungsi Transformasi fitur

Katakanlah Anda memiliki beberapa poin dari 2 jenis dalam sebuah makalah yang linear dipisahkan. *SVM* akan menemukan garis lurus yang memisahkan titik-titik menjadi 2 jenis dan terletak sejauh mungkin dari semua titik tersebut. Dalam hal skala,

beberapa masalah terbesar yang telah dipecahkan menggunakan *SVM* (dengan implementasi sesuai dimodifikasi) adalah iklan display, rekognasi situs sambatan manusia, deteksi jenis kelamin berbasis citra, *klasifikasi citra* berskala besar. *Support vector machine (SVM)* membentuk *hyperplanes* dan pada dasarnya digunakan untuk *klasifikasi dan analisis regresi* dan telah digunakan untuk banyak aplikasi. Ini adalah model *supervise learning* dan *pengklasifikasi biner* yang membuat prediksi berdasarkan data yang dirasakan sebelumnya. Dalam *SVM*, semua titik data adalah milik satu kelas atau kelas lain dan tujuannya adalah untuk memutuskan kelas mana yang cocok untuk titik data baru. Semua titik data yang berada di satu sisi *hyperplane* diberi label sebagai P dan yang lainnya adalah P - 1. Titik data yang membantu untuk menentukan hyperplane dan dengan demikian terdekat dengan hyperplane disebut vektor dukungan. (Patel, 2018)

Berdasarkan asumsi diatas berdasarkan asumsi bahwa kedua belah kelas terpisah secara sempurna oleh *hyperplane*. Dua buah kelas tidak selalu terpisah secara sempurna. Hal tersebut menyebabkan *constraint* pada persamaan 1 tidak terpenuhi, sehingga optimasi tidak terpenuhi dilakukan. Untuk mengatasi masalah ini, *SVM* dirumuskan ulang dengan menggunakan teknik *soft margin*. *Soft margin* dijelaskan dalam bentuk gambar II-1. Soft margin digunakan untuk kasus dimana data tidak liner untuk dipisahkan. (Sembiring, 2007)



Gambar II-1 Soft margin dalam *SVM (Sembiring, 2007)*

Pada umumnya permasalahan data tidak dapat dipisahkan secara *Linear* dalam ruang *input*, *soft margin SVM* tidak dapat menemukan pemisah dalam *hyperplane* sehingga tidak dapat memiliki akurasi yang besar dan tidak menggeneralisasi dengan baik. Oleh karena itu, dibutuhkan *kernel* untuk mentransformasikan data ke ruang dimensi yang lebih tinggi yang disebut ruang *kernel* yang berguna untuk memisahkan data secara *Linear*. Secara umum, fungsi *kernel* yang sering digunakan adalah kernel *Linear*, *Polynomial* dan *Radial Basis Function* (RBF). (Samsudiney, 2019)

Menurut (Sembiring, 2007) Untuk mengklasifikasikan data yang tidak dapat dipisahkan secara *linier formula SVM* harus dimodifikasi karena tidak akan ada solusi yang ditemukan. Oleh karena itu, kedua bidang pembatas harus diubah sehingga lebih *fleksibel* (untuk kondisi tertentu) dengan penambahan *variabel* ξ i (i ≥ ∀i ξ 0, : ξ i = 0 jika i x diklasifikasikan dengan benar) menjadi xi .w + b ≥1-ξi untuk kelas 1 dan xi .w + b ≤ −1+ξi untuk kelas 2. Pencarian bidang pemisah terbaik dengan dengan penambahan *variabel* ξ i sering juga disebut *soft margin hyperplane.* Dengan demikian formula pencarian bidang pemisah terbaik berubah menjadi:

$$\min \frac{1}{2}|w|^2 + C\left(\sum_{i=1}^{n} \xi_i\right)$$
$$s.t. y_i (w.x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

(2.2)

C adalah parameter yang menentukan besar penalti akibat kesalahan dalam klasifikasi data dan nilainya ditentukan oleh pengguna.

Pada penelitian yang dilakukan oleh (Jacot, Gabriel and Hongler, 2018) *SVM* dapat dikombinasikan dengan *NTK* dengan hasil yang sangat memuaskan.

2.1.9 Neural Tangen Kernel (*NTK*)

Neural Tangent Kernel merupakan sebuah penemuan pada tahun 2018 dimana fungsi utamanya adalah dapat mempelajari jaringan saraf tiruan yang sangat rumit menjadi lebih simple dengan membuatnya kedalam sebuah kernel terlebih dahulu.

Secara garis besar *NTK* dapat didefinisikan :

- Kernel yang berfungsi agar membuat orang memahami ANN dengan mempelajari kernelnya.
- Evolusi dari ANN sewaktu di training dapat digambarkan dengan kernel *NTK*.
- Kernel yang dibuat lebih simple.

Jaringan saraf tiruan (JST) telah mencapai hasil yang mengesankan di berbagai bidang *machine learning*. Padahal telah lama diketahui bahwa *JST* dapat mendekati fungsi apapun dengan cukup banyak neuron tersembunyi , tidak diketahui *konvergensi pengoptimalan JST*. Ciri misterius lain dari *JST* adalah sifat *generalisasinya* yang baik meskipun biasanya *over-parametrization*. Tampaknya paradoks bahwa jaringan saraf yang cukup besar dapat masuk secara acak Pracetak. *Kernel* digunakan dalam *inferensi Bayesian* atau *Mesin Vektor Dukungan,* memberikan hasil yang sebanding dengan *JST* dilatih dengan penurunan *gradien*. Kita akan melihat bahwa dalam batas yang sama, perilaku *JST* selama pelatihan dijelaskan oleh kernel terkait, yang disebut *Neural Tangen Kernel (NTK).* (Jacot, Gabriel and Hongler, 2018)

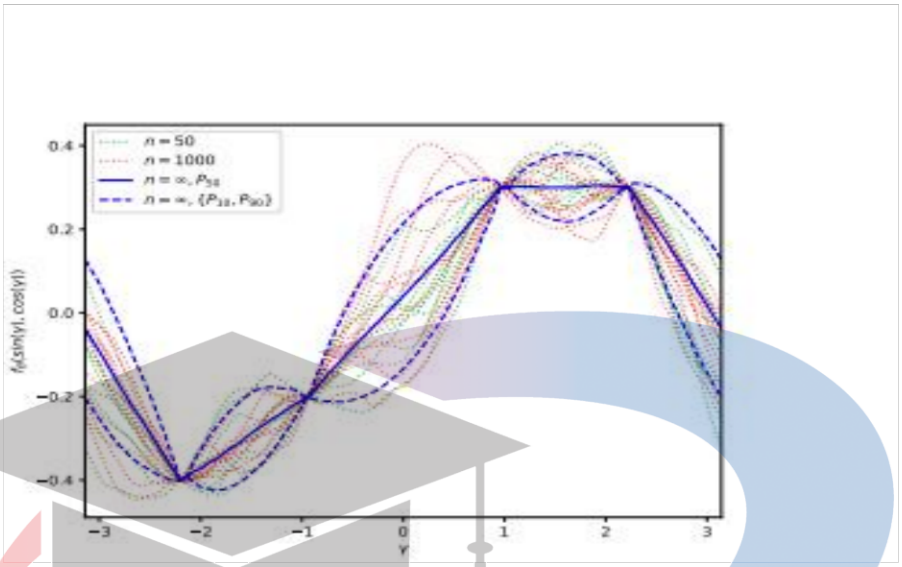
NTK adalah *kernel* yang dicirikan oleh turunan dari *output jaringan* ke parameternya. Itu telah ditunjukkan bahwa *NTK* dari jaringan dengan *inisialisasi Gaussian* bertemu dengan *deterministik kernel* dan tetap tidak berubah selama *penurunan gradien* dalam batas lebar tak terhingga. *NTK* memperpanjang ini hasil untuk kasus *inisialisasi ortogonal* dan menemukan bahwa *bobot ortogonal* berkontribusi sama *properti* untuk *NTK*. Dengan kecepatan pembelajaran yang cukup kecil dan lebar yang lebar, jaringan dioptimalkan oleh penurunan gradien berperilaku sebagai model yang *dilinierisasi* tentang *parameter* awalnya , di mana ini dinamikanya disebut *rezim NTK,* atau *lazy training* (Xu, Du and Huang, 2020).

Dalam studi *jaringan saraf tiruan (JST), Neural Tangent Kernel (NTK)* adalah *kernel* yang menjelaskan *evolusi jaringan saraf tiruan* dalam selama pelatihan mereka dengan penurunan gradien. Hal ini memungkinkan *JST* untuk dipelajari menggunakan alat teoritis dari *Metode Kernel*. *Neural Tangent Kernel* adalah sebuah kernel didefinisikan oleh

$$f(\cdot,\theta):\mathbb{R}^{n_{\text{in}}}\to\mathbb{R}$$

$$\Theta(x,y;\theta)=\sum_{p=1}^{P}\partial_{\theta_p}f(x;\theta)\,\partial_{\theta_p}f(y;\theta)\,. \qquad (2.3)$$

Untuk *arsitektur jaringan neural* yang paling umum, *NTK* menjadi konstan dalam batas lebar lapisan yang besar. Ini memungkinkan pembuatan pernyataan formulir tertutup sederhana tentang *prediksi jaringan neural, dinamika pelatihan, generalisasi,* dan *permukaan kerugian.* Misalnya, ini menjamin bahwa *JST* yang cukup lebar menyatu ke *minimum global* saat dilatih untuk meminimalkan *kerugian empiris.* (Jacot, Gabriel and Hongler, 2018) . Contoh dari gambar *NTK*



Gambar II-2 Contoh Gambar *NTK* (*(Jacot, Gabriel and Hongler, 2018)*

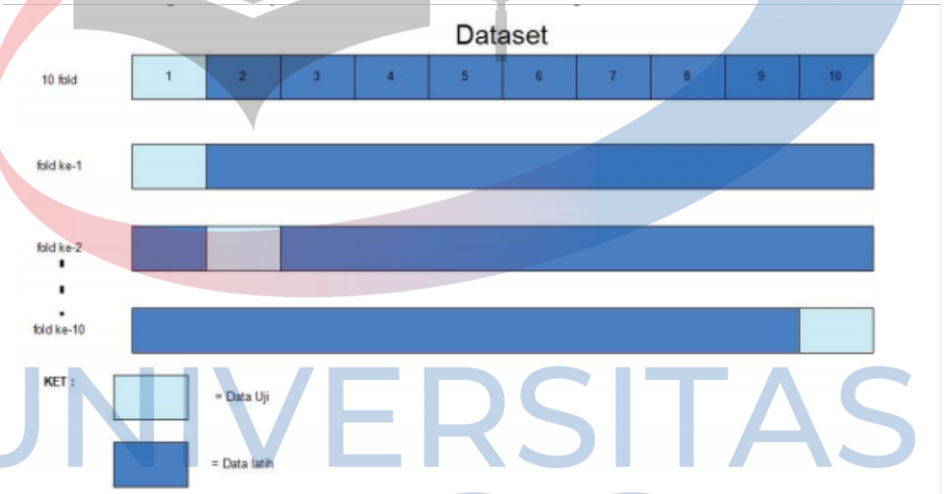
NTK juga menjadi dasar dari terbentuknya *GNTK (Graph Neural Tangent Kernels)*. Selain itu, fungsi kernel seringkali memiliki ekspresi eksplisit, dan dengan demikian kita dapat menganalisis jaminan teoretis mereka dengan menggunakan alat bantu dalam teori pembelajaran (Du *et al.*, 2019)

2.1.10 Cross-validation (CV)

Cross-validation (CV) adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua subset yaitu data proses pembelajaran dan data validasi / evaluasi. Model atau algoritma dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi. Selanjutnya pemilihan jenis CV dapat didasarkan pada ukuran *Dataset*. Biasanya CV K-fold digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi. K-fold cross validation adalah salah satu metode untuk mengevaluasi kinerja classifier, metode ini dapat digunakan apabila memiliki jumlah data yang terbatas (jumlah instance tidak banyak) . K-fold cross validation merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu

sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak. (Ghojogh and Crowley, 2019)

K-fold cross validation diawali dengan membagi data sejumlah n-fold yang diinginkan. Dalam proses cross validation data akan dibagi dalam n buah partisi dengan ukuran yang sama D1, D2, D3 .. Dn selanjutnya proses uji dan latih dilakukan sebanyak n kali. Dalam iterasi ke-i partisi Di akan menjadi data uji dan sisanya akan menjadi data latih. Untuk penggunaan jumlah fold terbaik untuk uji validitas, dianjurkan menggunakan 10-fold cross validation dalam model. Contoh pembagian *Dataset* dalam proses 10-fold cross validation terlihat pada Gambar II-12



Gambar II-12 Contoh iterasi data dengan k-fold cross validation (Sitorus, 2020)

Menurut (Nugroho, 2019) cara kerja K-fold cross validation adalah sebagai berikut:

1. Total instance dibagi menjadi N bagian.
2. Fold ke-1 adalah ketika bagian ke-1 menjadi data uji (testing data) dan sisanya menjadi data latih (training data). Selanjutnya, hitung akurasi berdasarkan porsi

data tersebut. Perhitungan akurasi tersebut dengan menggunakan persamaan sebagai berikut :

$$Akurasi = \frac{\sum data\ uji\ benar\ klasifikasi}{\sum total\ data\ uji} \times 100\% \quad (2.7)$$

3. Fold ke-2 adalah ketika bagian ke-2 menjadi data uji (testing data) dan sisanya menjadi data latih (training data). Selanjutnya, hitung akurasi berdasarkan porsi data tersebut.
4. Demikian seterusnya hingga mencapai fold ke-K. Hitung rata-rata akurasi dari K buah akurasi di atas. Rata-rata akurasi ini menjadi akurasi final.

2.1.11 Confusion Matrix

Percobaan dari penelitian dievaluasi dengan pengukuran akurasi, presisi, *Recall* dan fmeasure. Pengukuran dilakukan dengan menggunakan tabel klasifikasi yang bersifat prediktif, disebut juga dengan Confusion Matrix (Xhemali, et al. 2009).

		True Values	
		True	False
Prediction	True	TP Correst result	FP Unexpected result
	False	FN Missing result	TN Correct absence of result

Gambar II-3 Tabel Confusion Matrix

Menurut (Rahmad, Suryanto and Ramli, 2020) dari Confusion Matrix dapat diukur akurasi, presisi dan *Recall* untuk menganalisa kinerja dari algoritma dalam melakukan klasifikasi untuk mendeteksi penyakit. Akurasi merupakan persentase dari prediksi yang benar. Presisi adalah ukuran dari akurasi dari suatu kelas tertentu yang telah diprediksi. *Recall* merupakan persentase dari data dengan nilai positif yang nilai prediksinya juga positif. Adapun perhitungannya adalah sebagai berikut:

Akurasi = (TP+TN) / (TP+FP+TN+FN) (2.4)

Presisi = TN / (FP+TN) (2.5)

Recall = TP / (TP+FN) (2.6)

Menurut (Nugroho, 2019) berikut adalah beberapa manfaat dari confusion matrix:

1. Menunjukkan bagaimana model ketika membuat prediksi.
2. Tidak hanya memberi informasi tentang kesalahan yang dibuat oleh model tetapi juga jenis kesalahan yang dibuat.
3. Setiap kolom dari confusion matrix merepresentasikan instance dari kelas prediksi.
4. Setiap baris dari confusion matrix mewakili instance dari kelas aktual.

2.2 Penelitian Terdahulu

Dalam melakukan penelitian ini, penulis menggunakan penelitian dahulu sebagai referensi bahwa penelitian ini memang merupakan permasalahan yang diperlukan pemecahannya dalam dunia ilmu pengetahuan.

Penelitian terdahulu yang pernah dilakukan yang berkaitan dengan penelitian ini adalah :

1. Structured Output *SVM* Prediction of Apparent Age, Gender and Smile From Deep *Feature*s (Uricar *et al.*, 2016)
2. A Comparative Analysis of Techniques for Predicting Student Performance (Bydžovská, 2016)
3. Data Descriptor: Open University Learning Analytics *Dataset* (Kuzilek, Hlosta and Zdrahal, 2017)
4. *Neural Tangent Kernel*: Convergence and generalization in neural networks (Jacot, Gabriel and Hongler, 2018)
5. GritNet: Student Performance Prediction with Deep Learning (Kim, Vizitei and Ganapathi, 2018)

6. Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika (Putri and Waspada, 2018)
7. Supervised data mining approach for predicting student performance. (Yaacob *et al.*, 2019)
8. Data Mining Untuk Penentuan Model Tingkat Kesuksesan Kelulusan Murid SMA Pada Perguruan Tinggi Negeri : Studi Kasus Di Iain Bone (Irmawati, Zainuddin and Yuyun, 2020)
9. An overview and comparison of supervised data mining techniques for student exam performance prediction. (Tomasevic, Gvozdenovic and Vranes, 2020)
10. The potential for student performance prediction in small cohorts with minimal available attributes (Wakelam *et al.*, 2020)

Berikut penulis buatkan dalam bentuk tabel sehingga memudahkan pembaca untuk melihatnya.

Tabel II-1 Ringkasan Penelitian Terdahulu

No	Judul	Hasil
1.	Structured Output *SVM* Prediction of Apparent Age, Gender and Smile From Deep *Features* (Uricar *et al.*, 2016)	Penelitian ini berhasil baik memprediksi usia, jenis kelamin dengan menggunakan metode *SVM* yang dikombinasikan dengan Deep *Features*.
2.	A Comparative Analysis of Techniques for Predicting Student Performance (Bydžovská, 2016)	Penelitian ini menyimpulkan *SVM* merupakan metode terbaik untuk prediksi.

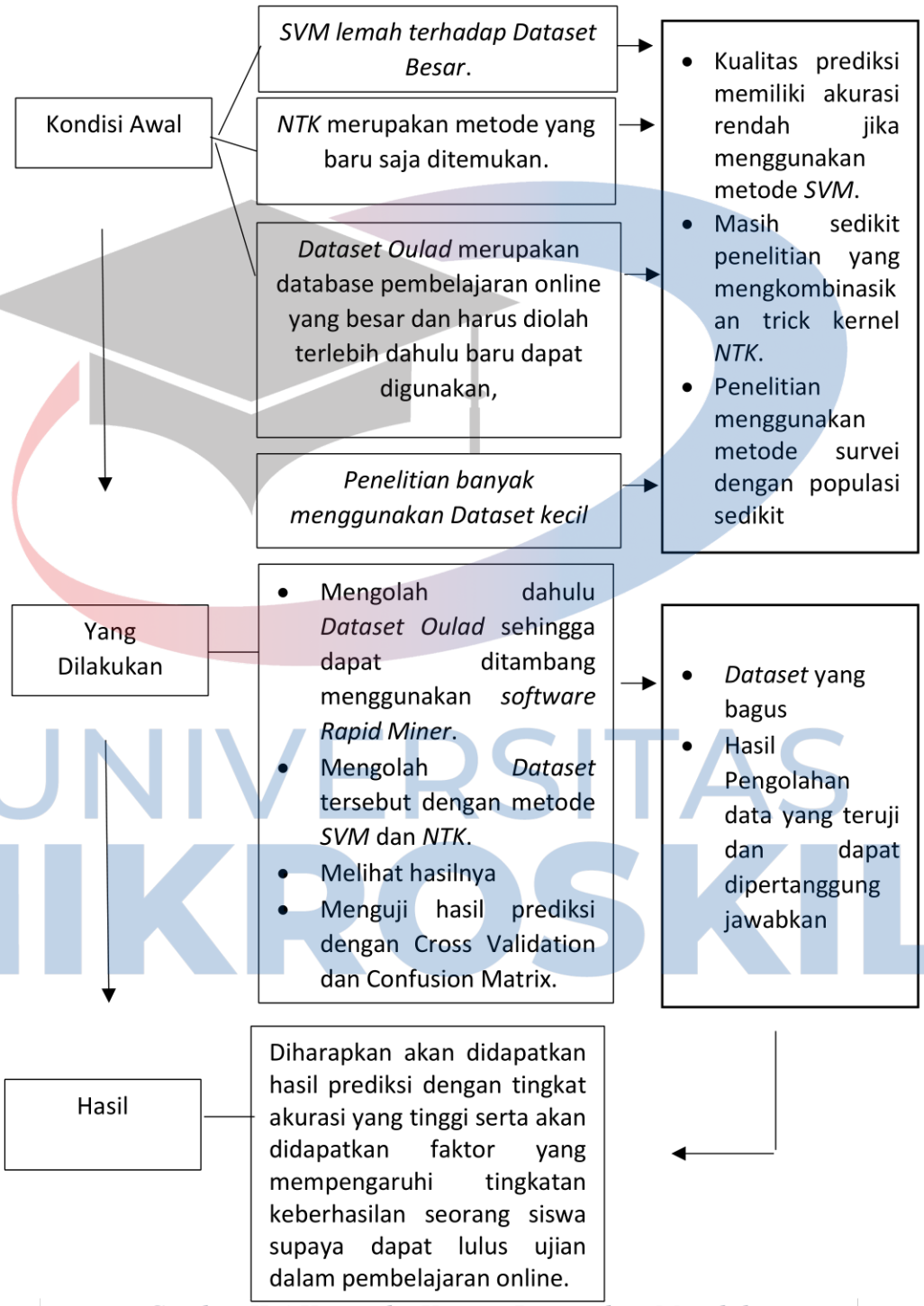
3.	Data Descriptor: Open University Learning Analytics *Dataset* (Kuzilek, Hlosta and Zdrahal, 2017)	Merupakan Penelitian tentang *Dataset Oulad* yang banyak digunakan oleh peneliti.
4.	*Neural Tangent Kernel*: Convergence and generalization in neural networks (Jacot, Gabriel and Hongler, 2018)	Merupakan Penemu dari *NTK*. Disini diuraikan penggunaan *NTK* yang dapat dikombinasikan dengan *SVM.*
5.	Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika (Putri and Waspada, 2018)	Algoritma C4.5 dapat digunakan untuk menghasilkan prediksi kelulusan dengan nilai rata-rata precision 63.93%, *Recall* 60.73%, dan akurasi 60.52%.
6.	GritNet: Student Performance Prediction with Deep Learning (Kim, Vizitei and Ganapathi, 2018)	Hasil menunjukkan GritNet tidak hanya secara konsisten mengungguli metode berbasis regresi logistik standar, tetapi peningkatan tersebut secara substansial diucapkan dalam beberapa minggu pertama ketika prediksi yang akurat paling menantang. Yang menarik adalah GritNet itu dibagun dari JST yang merupakan dasar dari *Neural Tangent Kernel* yang akan digunakan dalam penelitian ini.
7.	Supervised data mining approach for predicting student performance. (Yaacob *et al.*, 2019)	Menampilkan hasil dari *Supervised Data Mining* untuk memprediksi Performa murid berdasarkan hasil ujian.

8.	Data Mining Untuk Penentuan Model Tingkat Kesuksesan Kelulusan Murid SMA Pada Perguruan Tinggi Negeri : Studi Kasus Di Iain Bone (Irmawati, Zainuddin and Yuyun, 2020)	Penelitian ini mendapatkan Nilai Precision sebesar 100%, *Recall* 100%, Accuracy 100%.
9.	An overview and comparison of supervised data mining techniques for student exam performance prediction. (Tomasevic, Gvozdenovic and Vranes, 2020)	Didapat kan hasil yang cukup baik dengan metode *SVM* tetapi tidak pernah dikombinasikan dengan *NTK*.
10.	The potential for student performance prediction in small cohorts with minimal available attributes (Wakelam *et al.*, 2020)	Merupakan Prediksi performa Siswa menggunakan *Algoritma Data Mining* dengan Hasil dan *Dataset* yang tidak begitu memuaskan.

2.3 Kerangka Konsep/Pola Pikir Pemecahan Masalah

Kerangka konsep pemecahan masalah menggambarkan bagaimana masalah penelitian dapat diselesaikan melalui solusi-solusi yang diusulkan serta dari solusi tersebut diharapkan memiliki dampak yang dapat menyelesaikan permasalahan penelitian.

Berikut adalah gambaran kerangka konsep pemecahan masalah dari penelitian yang akan dilakukan, dapat dilihat pada Gambar II-4. berikut:



Gambar II-4 Kerangka Konsep Pemecahan Masalah

Metode *SVM* merupakan metode yang bagus untuk prediksi tetapi memiliki kelemahan dalam mengolah *Dataset* besar. Untuk itu diperlukan bantuan Trick kernel untuk mendapatkan hasil yang bagus. Dalam penelitian ini penulis memilih *Neural Tangent Kernel* (*NTK*). Hal ini berdasarkan rekomendasi dari penelitian yang dilakukan oleh (Jacot, Gabriel and Hongler, 2018).

Sangat sedikitnya penelitian yang menggunakan *NTK* karena merupakan metode yang baru saja ditemukan. Pada hal menurut penelitian Metode *NTK* cocok untuk mengatasi kekurangan *SVM* dalam mengolah data yang besar. *Dataset Oulad* merupakan database yang besar mempunyai banyak atribut dan harus diolah terlebih dahulu baru dapat digunakan, membuat hampir tidak ada peneliti Indonesia yang menggunakannya. Mengingat *Dataset* ini juga merupakan *Dataset* pembelajaran online yang masih dapat dibilang baru. Dimana biasanya di Indonesia menggunakan *Dataset* dengan populasi kecil karena pembelajaran tatap muka offline yang hanya dapat memfalitasi sedikit siswa dan sulit dalam pendokumentasiannya dibanding pembelajaran online.

Mengolah dahulu *Dataset Oulad* sehingga dapat ditambang menggunakan *software Rapid Miner*. Pada tahap ini dilakukan dahulu *Data Preprocessing* (Data Cleaning, *Feature* Selection, dan pembagian *Dataset* untuk training dan testing dengan metode Train-Test Split. Dalam proses preprocessing akan banyak menggunakan software *Microsoft Excel*. *Microsoft Excel* adalah software pengolah angka dari Microsoft yang dijual dalam pake Microsoft Office.

Mengolah *Dataset yang sudah dipreprocessing* tersebut dengan metode *SVM* dan *NTK*. Hasil kemudian akan diuji dengan metode cross validation dan dianalisis dengan metode confusion matrix.

Diharapkan akan didapatkan hasil dengan tingkat akurasi yang tinggi serta akan didapatkan faktor yang mempengaruhi tingkatan keberhasilan seorang siswa supaya dapat lulus ujian.

Pada penelitian ini, penulis menggunakan software RapidMiner dan program pendukung lainnya seperti *Phyton* untuk mengelola *Dataset Oulad*. *Dataset Oulad*

merupakan suatu *Dataset* besar sehingga peneliti harus memprosesnya terlebih dahulu sebelum dapat digunakan di software *Rapid Miner*. Penulis akan mengelolanya menggunakan Komputer dengan spesifikasi Core i-7 , memory DDR3 8 GB.

Jika terdapat hal yang diperlukan dalam penelitian dimana data tersebut tidak dapat diolah dengan rapidminer penulis akan membuat programnya dengan menggunakan bahasa pemograman *Phyton.* Saat ini, *Python* juga merupakan bahasa yang populer bagi bidang *data science* dan analisis. Hal ini dikarenakan oleh dukungan bahasa Python terhadap library – library yang didalamnya menyediakan fungsi analisis data dan fungsi *machine learning, Data Preprocessing tools*, serta *visualisasi data*. Untuk menganalisis data apalagi data yang cukup besar diperlukan suatu alat untuk menghitung dan menganalisis supaya lebih efektif dan efisien. *Python* merupakan salah satu alat yang direkomendasikan untuk hal tersebut. Dalam beberapa tahun terakhir, Dukungan perpustakaan *Python* yang ditingkatkan (terutama panda) telah membuatnya menjadi alternatif yang kuat untuk tugas analisis data (Sodik, Dwi and Kharisudin, 2020) .

Secara umum, menurut (Edwardo, 2018) *Python* memiliki ciri-ciri sebagai berikut:

1. Banyak mendukung *library*
2. Bahasa yang relatif mudah dipahami
3. Memiliki aturan *layout source code* yang memudahkan pengecekan *code*
4. Bahasa yang *interpreted* karena *code* dieksekusi satu per satu dan melakukan *debugging* lebih mudah dibandingkan dengan bahasa yang di-compile
5. Bahasa yang *portable* karena hanya *code* satu kali untuk menjalankan di *platform* lain
6. Bahasa yang *open-source*.

Berikut ini menurut (Edwardo, 2018) adalah beberapa alasan *Python* menjadi bahasa yang populer, khususnya dalam *ranah analisis data* dan *data science* :

1. Ketersediaan akan *open-source library, frameworks, tools untuk data mining*, contohnya adalah *SciKit Learn, TensorFlow, Keras.*

2. Relatif lebih mudah dipahami. Penulisan code di *Python* relatif lebih singkat dibandingkan bahasa pemrograman yang lain.
3. Multifungsi, tidak hanya untuk data processing, namun juga bisa untuk tugas lain seperti membuat website dan tampilan *GUI (Graphical User Interface)*.

2.4 Hipotesis

Berdasarkan permasalahan yang telah diajukan pada bagian sebelumnya. Maka dapat ditarik suatu hipotesis atas permasalahan tersebut. Hipotesis tersebut adalah sebagai berikut :

1. Dengan menggunakan metode *SVM* dikombinasikan dengan *NTK* maka dapat menyelesaikan / menambahkan suatu sumbangan yang berdampak besar bagi ilmu pengetahuan ke penelitian sebelumnya.
2. Penggunaan *Dataset* yang besar dan berkwalitas internasional diharapkan dapat menjadi nilai lebih dari penelitian yang dilakukan ini . Mengingat *Dataset* yang digunakan banyak juga digunakan oleh penelitian yang dipublikasi di jurnal bereputasi.