|  |  |
| --- | --- |
|  |  |
|  | **MODUL 5 PERKULIAHAN** |
|  |  |
|  | **Proyek Data Science** |
|  |  |
|  | **Studi kasus 2 penerapan algoritma data science** |
|  |  |

# Pembahasan

**Pertemuan 5**

**Studi kasus 2 penerapan algoritma data science**

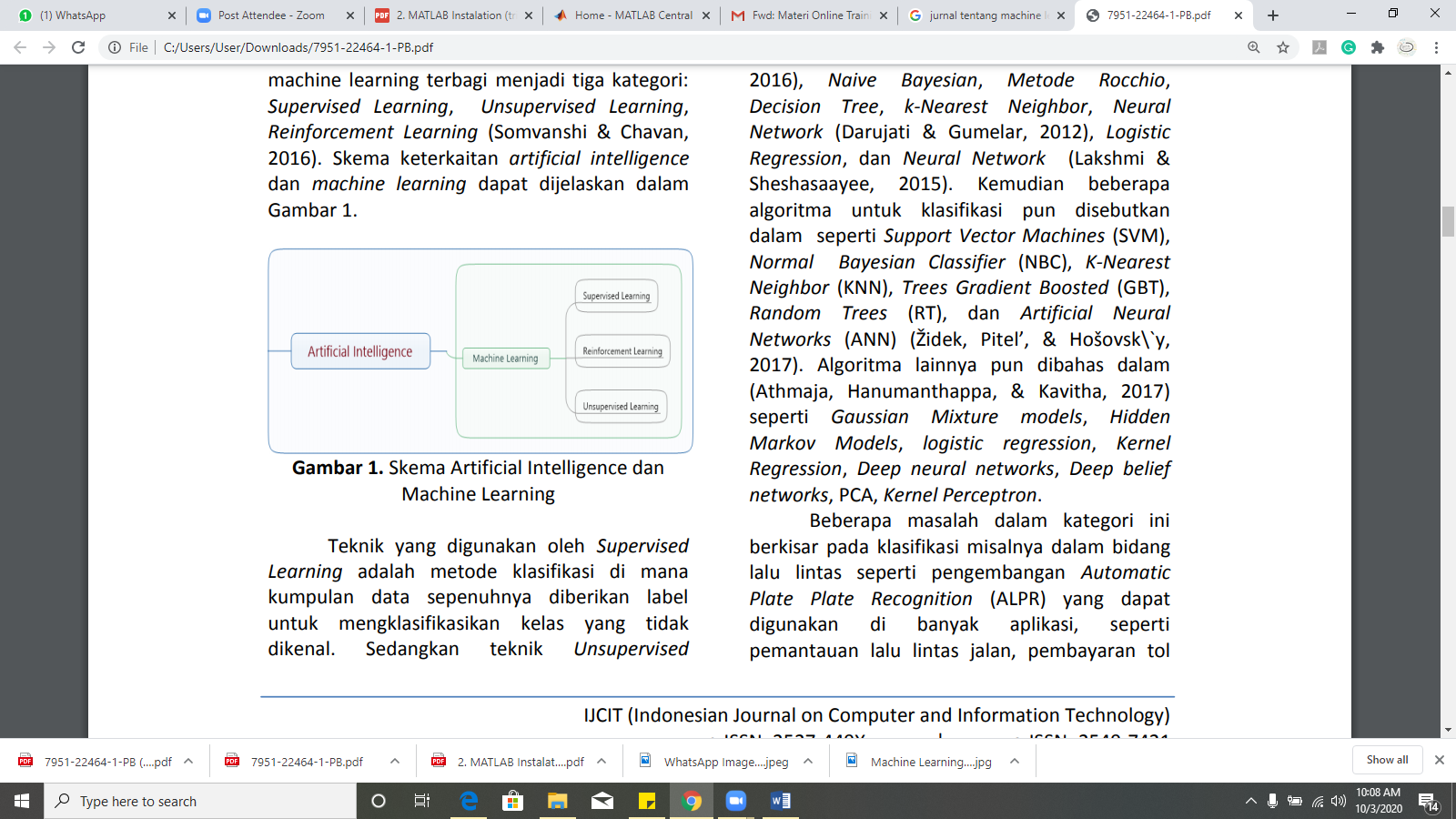
Berdasarkan jurnal : Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper

Sumber : Ahmad Roihan1 , Po Abas Sunarya2 , Ageng Setiani Rafika3 1,3Sistem Komputer, Universitas Raharja Tangerang, Indonesia e-mail: ahmad.roihan@raharja.info1 , agengsetianirafika@raharja.info3 2 Teknik Informatika, Universitas Raharja Tangerang, Indonesia e-mail: abas@raharja.info

Secara bertahap penelitian tentang penggunaan perangkat sensor semakin berkembang. Beberapa diantaranya digunakan dalam proses monitoring (Ahmad Roihan, Permana, & Mila, 2016), alat ukur (Supriyono, Sudarto, & Fakhri, 2016), simulator (A. Roihan, Sunarya, & Wijaya, 2019), dan lain sebagainya. Namun dalam penggunaannya tersebut masih perlu pengguna sebagai pengendali sistem dan belum menunjukkan alat dengan fungsi yang cerdas. Seiring berlalunya waktu, mesin pintar atau cerdas perlahan-lahan akan menggantikan dan meningkatkan kemampuan manusia di berbagai bidang (Nayak & Dutta, 2017). Kecerdasan yang ditunjukkan oleh mesin biasanya disebut dengan kecerdasan buatan (Artificial Intelligence) yang merupakan bagian dari ilmu komputer. Kecerdasan Buatan merupakan salah satu bidang dalam ilmu komputer yang ditujukan pada pembuatan software dan hardware yang dapat berfungsi sebagai sesuatu yang dapat berpikir seperti manusia (Sunarya, Santoso, & Sentanu, 2015).

Kecerdasan buatan banyak digunakan untuk memecahkan berbagai masalah seperti bisnis (Rahardja, Roihan, & others, 2017), robotika, bahasa alami, matematika, game, persepsi, diagnosis medis, teknik, analisis keuangan, analisis sains, dan penalaran (Russell & Norvig, 2016). Machine learning dapat didefinisikan sebagai aplikasi komputer dan algoritma matematika yang diadopsi dengan cara pembelajaran yang berasal dari data dan menghasilkan prediksi di masa yang akan datang (Goldberg & Holland, 1988). Adapun proses pembelajaran yang dimaksud adalah suatu usaha dalam memperoleh kecerdasan yang melalui dua tahap antara lain latihan (training) dan pengujian (testing) (Huang, Zhu, & Siew, 2006).

Bidang machine learning berkaitan dengan pertanyaan tentang bagaimana membangun program komputer agar meningkat secara otomatis dengan berdasar dari pengalaman (Mitchell, 1997). Penelitian terkini mengungkapkan bahwa machine learning terbagi menjadi tiga kategori: Supervised Learning, Unsupervised Learning, Reinforcement Learning (Somvanshi & Chavan, 2016). Skema keterkaitan artificial intelligence dan machine learning dapat dijelaskan dalam Gambar 1.



Gambar 1. Skema Artificial Intelligence dan Machine Learning

Teknik yang digunakan oleh Supervised Learning adalah metode klasifikasi di mana kumpulan data sepenuhnya diberikan label untuk mengklasifikasikan kelas yang tidak dikenal. Sedangkan teknik Unsupervised Learning sering disebut cluster dikarenakan tidak ada kebutuhan untuk pemberian label dalam kumpulan data dan hasilnya tidak mengidentifikasi contoh di kelas yang telah ditentukan (Thupae, Isong, Gasela, & AbuMahfouz, 2018). Sedangkan Reinforcement Learning biasanya berada antara Supervised Learning dan Unsupervised Learning (Board, 2017), teknik ini bekerja dalam lingkungan yang dinamis di mana konsepnya harus menyelesaikan tujuan tanpa adanya pemberitahuan dari komputer secara eksplisit jika tujuan tersebut telah tercapai (Das & Nene, 2017).

Metode supervised learning didasarkan pada kumpulan sampel data yang memiliki label. Kumpulan sampel digunakan untuk meringkas karakteristik distribusi ukuran perilaku dalam setiap jenis aplikasi sehingga membentuk model perilaku dari data (Amei, Huailin, Qingfeng, & Ling, 2011). Supervised learning dikelompokkan lebih lanjut dalam masalah klasifikasi dan regresi. Masalah klasifikasi adalah ketika variabel output berbentuk kategori, seperti merah atau biru atau penyakit dan tidak ada penyakit. Sedangkan masalah regresi adalah ketika variabel output adalah nilai riil, seperti dollar atau berat (Brownlee, 2016). Supervised learning memiliki beberapa algoritma populer seperti Back-propagation (Negnevitsky, 2005), Linear regression, Random Forest, Support Vector Machines (Brownlee, 2016), Naive Bayesian, Metode Rocchio, Decision Tree, k-Nearest Neighbor, Neural Network (Darujati & Gumelar, 2012), Logistic Regression, dan Neural Network (Lakshmi & Sheshasaayee, 2015).

Kemudian beberapa algoritma untuk klasifikasi pun disebutkan dalam seperti Support Vector Machines (SVM), Normal Bayesian Classifier (NBC), K-Nearest Neighbor (KNN), Trees Gradient Boosted (GBT), Random Trees (RT), dan Artificial Neural Networks (ANN) (Židek, Pitel’, & Hošovsk\`y, 2017). Algoritma lainnya pun dibahas dalam (Athmaja, Hanumanthappa, & Kavitha, 2017) seperti Gaussian Mixture models, Hidden Markov Models, logistic regression, Kernel Regression, Deep neural networks, Deep belief networks, PCA, Kernel Perceptron. Beberapa masalah dalam kategori ini berkisar pada klasifikasi misalnya dalam bidang lalu lintas seperti pengembangan Automatic Plate Plate Recognition (ALPR) yang dapat digunakan di banyak aplikasi, seperti pemantauan lalu lintas jalan, pembayaran tol otomatis, dan manajemen parkir (Kosala, Harjoko, & Hartati, 2017).

Pemanfaatan machine learning pun dalam bidang industri dilakukan pada penelitian (Krisandi, Helmi, & others, 2013). Selain itu dalam bidang kedokteran seperti masalah pencitraan medis (Latif, Xiao, Imran, & Tu, 2019), pengelolaan data pasien (Khan, Doucette, Cohen, & Lizotte, 2012) dan pemeriksaan gaya berjalan seseorang bisa diprediksi dengan akurasi yang cukup tinggi (LeMoyne, Kerr, Mastroianni, & Hessel, 2014). Dalam bidang teknologi seperti jam pintar (Lee, Yoon, & Lee, 2018), multimedia (Holder, Pin, & Kalva, 2009), klasifikasi teks (Darujati & Gumelar, 2012) dan lain sebagainya tentu saja sudah banyak yang memanfaatkan machine learning dengan kategori supervised learning.

Dalam jenis pembelajaran Unsupervised Learning, sistem disediakan dengan beberapa input sampel tetapi tidak ada output yang hadir. Karena tidak ada output yang diinginkan di sini kategorisasi dilakukan sehingga algoritma membedakan dengan benar antara kumpulan data. Ini adalah tugas mendefinisikan fungsi untuk menggambarkan struktur yang tersembunyi dari data yang tidak berlabel (Somvanshi & Chavan, 2016). Unsupervised learning dikelompokkan lebih lanjut dalam masalah clustering dan asosiasi.

Masalah pengelompokan (clustering) adalah tempat untuk menemukan pengelompokan yang melekat dalam data, seperti mengelompokkan pelanggan berdasarkan pada perilaku pembelian. Sedangkan masalah asosiasi adalah aturan yang menggambarkan sebagian besar data yang ada, seperti orang yang membeli A juga cenderung membeli B (Brownlee, 2016). Unsupervised learning memiliki beberapa algoritma populer seperti k-means, Apriori (Brownlee, 2016), Independent Subspace Analysis (ISA) (G. Wu et al., 2013), DBSCAN (Ester, Kriegel, Sander, Xu, & others, 1996). Beberapa masalah misalnya dalam bidang finansial untuk meninjau sejumlah besar data maka unsupervised learning biasanya dapat digunakan (Board, 2017).

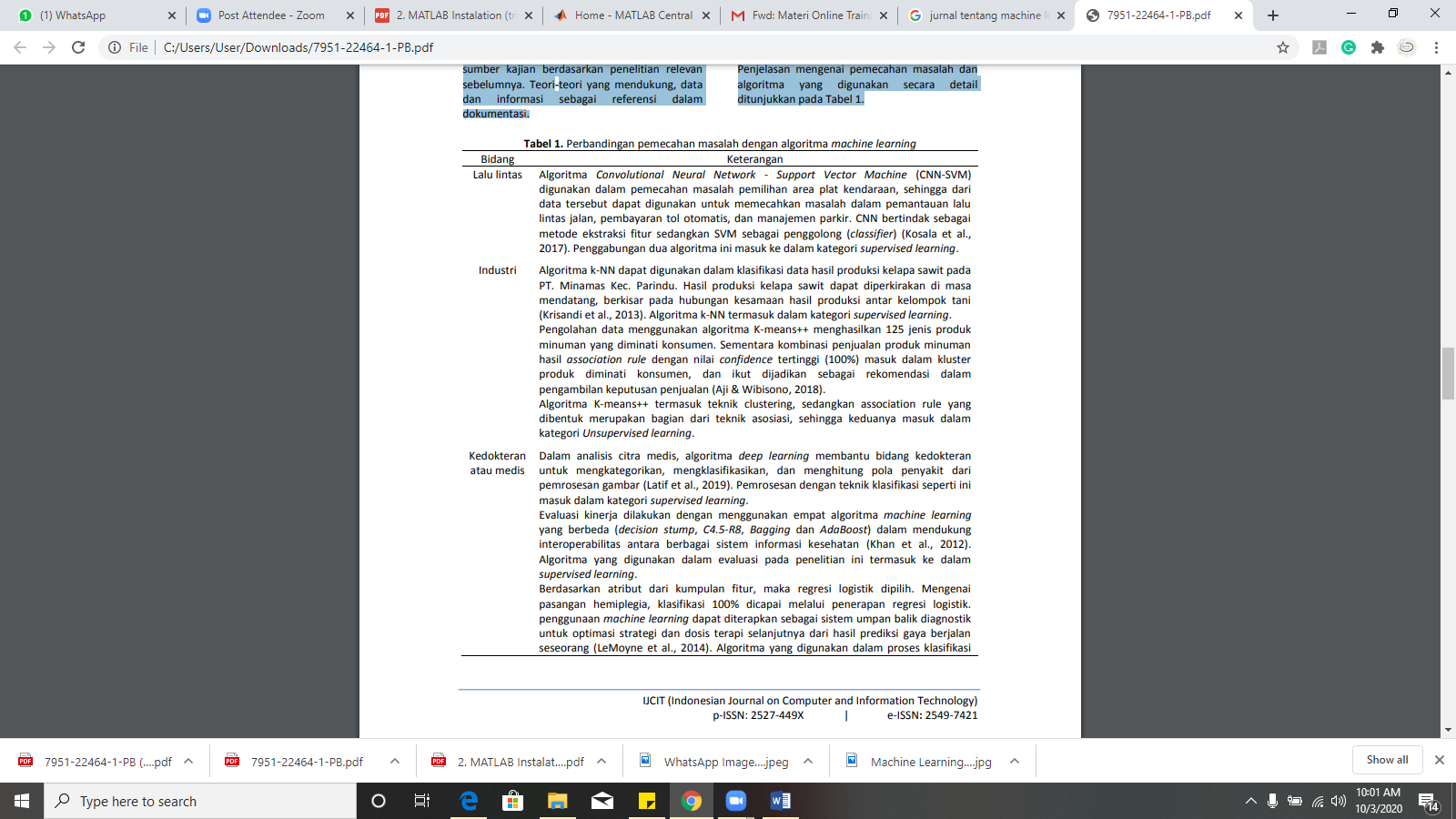
Dalam bidang industri misalnya dalam (Aji & Wibisono, 2018), kemudian dalam bidang kedokteran unsupervised learning digunakan dalam proses segmentasi pembuluh darah (Dharmawan, Ng, & Rahardja, 2018), dan bidang teknologi seperti jaringan komputer maupun pencegahan serangan keamanannya (Das & Nene, 2017) pun menggunakan kategori ini. Reinforcement learning berasal dari teori belajar hewan. Pembelajaran ini tidak memerlukan pengetahuan sebelumnya, dapat secara mandiri mendapatkan kebijakan opsional dengan pengetahuan yang diperoleh melalui coba-coba dan terus berinteraksi dengan lingkungan yang dinamis (Qiang & Zhongli, 2011). Masalah reinforcement learning diselesaikan dengan mempelajari pengalaman baru melalui trial-and-error (Mahmud, Kaiser, Hussain, & Vassanelli, 2018).

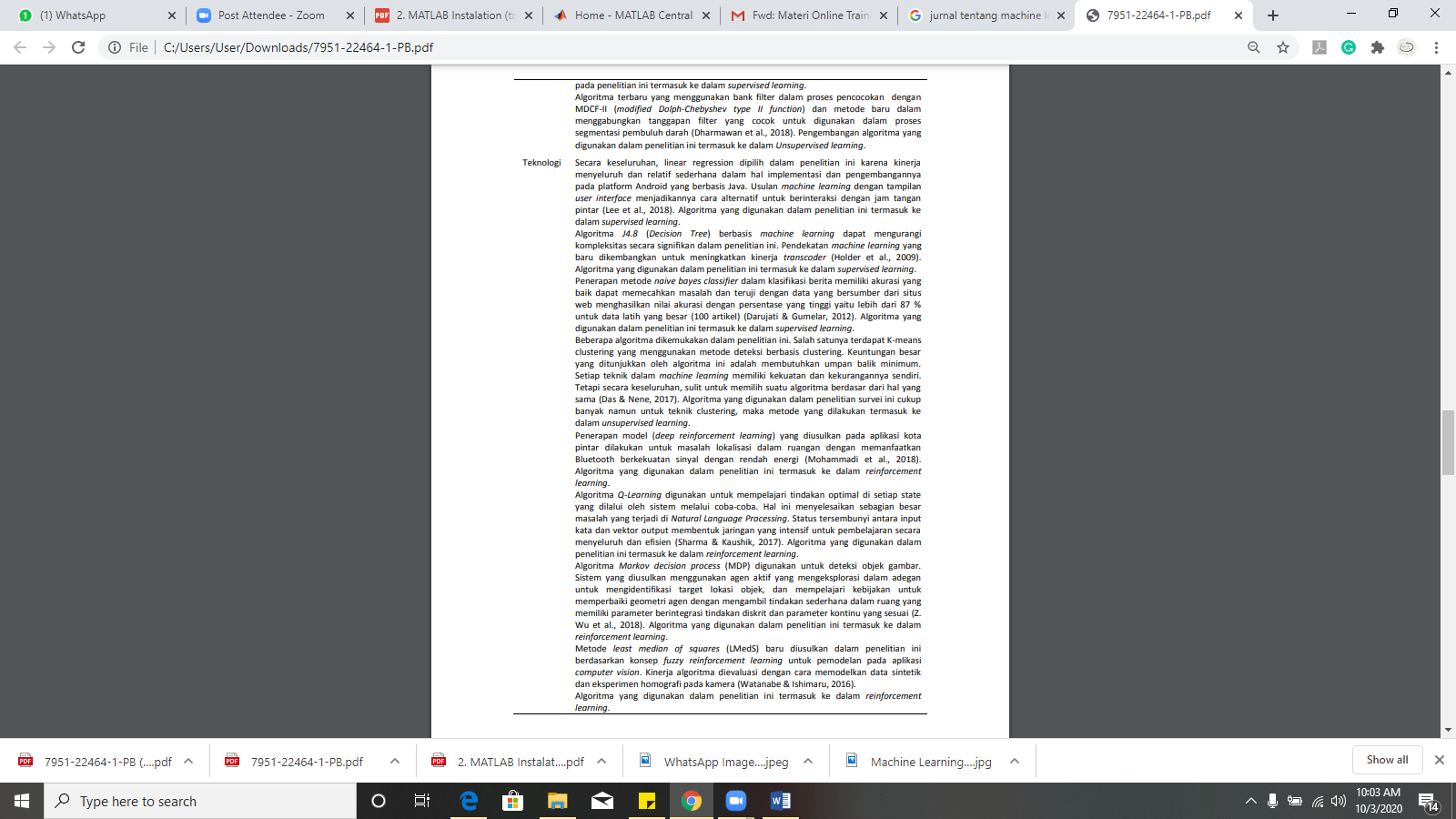
Algoritma reinforcement learning terkait dengan algoritma pemrograman dinamis yang sering digunakan untuk menyelesaikan masalah optimisasi (Mitchell, 1997). Banyak masalah machine learning dalam dunia nyata termasuk dalam kategori ini. Hal ini dikarenakan bisa mencapai harga yang mahal bahkan memakan waktu untuk pemberian label pada data terkait kemungkinan untuk memerlukan akses ke bagian pakar.

Padahal data yang tidak memiliki label itu didapatkan dengan harga yang murah dan mudah untuk dikumpulkan dan disimpan (Brownlee, 2016). Secara khusus metode reinforcement learning berdasarkan model proses pengambilan keputusan Markov (Markov decision process) mencakup dua jenis. Pertama yaitu metode berbasis model seperti algoritma SARSA, di mana reinforcement learning pertama kali mempelajari pengetahuan model, kemudian mendapatkan strategi yang optimal dari pengetahuan model tersebut. Kedua adalah metode yang relevan model seperti algoritma Temporal Difference dan algoritma Q-learning, di mana reinforcement learning secara langsung menghitung strategi yang optimal tanpa pengetahuan model (Qiang & Zhongli, 2011).

Beberapa masalah dalam kategori ini misalnya dalam bidang teknologi seperti pelayananan yang cerdas dengan memanfaatkan bidang IoT dan kota pintar (smart city) (Mohammadi, Al-Fuqaha, Guizani, & Oh, 2018). Dalam bidang teknologi lainnya seperti pemrosesan bahasa alami (Sharma & Kaushik, 2017) dan computer vision (Z. Wu, Khan, Gao, & Guan, 2018) dan (Watanabe & Ishimaru, 2016) juga menggunakan kategori ini. Artikel ini akan menyajikan beberapa penelitian terkini mengenai pemanfaatan algoritma machine learning berdasarkan masing-masing kategori, dengan harapan dapat menemukan celah dan dijadikan pedoman untuk penelitian pada masa yang akan dating. 2. METODE PENELITIAN Penelitian ini merupakan kajian pustaka dari beberapa artikel terkait machine learning.

Peninjauan dilakukan dari beberapa upaya penelitian terbaru yang memanfaatkan machine learning. Selanjutnya kajian ini berasal dari beberapa literasi dan mencakup upaya pemecahan masalah yang dibagi ke dalam bagian bidang-bidang dari perspektif masingmasing kategori machine learning. Proses pengumpulan data yang digunakan untuk memeriksa beberapa literatur sangat berguna untuk mencari dan memperoleh sumber kajian berdasarkan penelitian relevan sebelumnya. Teori-teori yang mendukung, data dan informasi sebagai referensi dalam dokumentasi. Ada empat bidang yang disajikan dalam artikel ini dari kajian beberapa artikel ilmiah. Bidang-bidang tersebut adalah lalu lintas, industri, kedokteran atau medis, dan teknologi. Semua bidang ini menerangkan teknik pembelajaran apa saja yang telah digunakan oleh beberapa peneliti. 3. HASIL DAN PEMBAHASAN Hasil dari ulasan menunjukkan beberapa penelitian menggunakan algoritma machine learning untuk memecahkan masalah sesuai dengan kebutuhan pada bidang masing-masing. Penjelasan mengenai pemecahan masalah dan algoritma yang digunakan secara detail ditunjukkan pada Tabel 1.





Dengan adanya kecerdasan buatan, diharapkan tidak menutup kemungkinan hanya dengan data pengetahuan yang terbatas, sebuah komputer dapat berpikir seperti manusia dalam menghadapi masalah (Sutrisno, Kristiadi, & Supriyanti, 2017). Tidak semua teknik bisa digunakan dalam segala bidang, namun semua algoritma yang ada dapat dikembangkan sehingga menjadi keterbaruan pengetahuan (novelty). Algoritma machine learning pun terus dikembangkan oleh beberapa penelitian. Tren terkini menunjukkan bahwa algoritma machine learning banyak dikembangkan dalam bidang kedokteran atau medis seperti yang dibahas dalam (Dharmawan, Li, Ng, & Rahardja, 2019), (Good et al., 2019), dan (Hinton, 2018).

4. KESIMPULAN Machine learning merupakan sub dari bidang keilmuan kecerdasan buatan (Artificial intelligence) yang banyak diteliti dan digunakan untuk memecahkan berbagai masalah. Ulasan dari berbagai bidang disajikan dalam bentuk pemecahan masalah dan algoritmanya dan dibagi menjadi tiga kategori dalam machine learning antara lain supervised learning, unsupervised learning, dan reinforcement learning. Ulasan dibatasi hanya pada beberapa bidang dan hasilnya menunjukkan bidang yang paling dominan terkini adalah bidang kedokteran atau medis diantara beberapa bidang lain seperti industri, teknologi dan lalu lintas. Penelitian lanjutan dapat dianjurkan untuk mengembangkan algoritma yang ada dan menerapkannya dalam aplikasi baik berbasis desktop maupun web (Židek et al., 2017), sehingga dapat dirasakan hasilnya secara langsung oleh pengguna akhir (end user).

Jurnal 2 dengan judul : PENERAPAN PENDEKATAN MACHINE LEARNING PADA PENGEMBANGAN BASIS DATA HERBAL SEBAGAI SUMBER INFORMASI KANDIDAT OBAT KANKER THE APPLICATION OF MACHINE LEARNING APPROACH TOWARDS HERBAL DATABASE DEVELOPMENT AS INFORMATION SOURCE FOR CANCER DRUG CANDIDATE Arli Aditya Parikesit\*), Rizky Nurdiansyah, dan David Agustriawan Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jl. Pulomas Barat Kav 88 Jakarta 13210 Email: [arli.parikesit@i3l.ac.id](mailto:arli.parikesit@i3l.ac.id)

Kanker merupakan penyakit yang mematikan yang disebabkan oleh proliferasi abnormal sel-sel tubuh. Sel-sel abnormal tersebut dapat merusak fungsi organ dan menyebabkan kematian. Insiden tahunan kanker diseluruh dunia kurang lebih sekitar 30,000 dimana jenis kanker yang paling mematikan adalah kanker paru-paru dan kanker payudara (Lim, 2002). Dengan meningkatnya prevalansi kanker, pengobatan komplementer dan alternatif sama-sama dibutuhkan sebagai obat kanker (Chrystal et al., 2003). Obat herbal dan tanaman seperti buah-buahan, sayur-sayuran, dan spesies tumbuhan lainnya memiliki banyak khasiat pada kesehatan (Lampe, 1999; Surh, 2003). Saat ini operasi, radioterapi dan kemotrapi merupakan pilihan utama untuk pengobatan kanker, sedangkan bahan senyawa alam masih dianggap sebagai tambahan pengobatan (Golbeck et al., 2011). Uniknya, 60 sampai 75% dari obat anti kanker merupakan produk dari senyawa alam. Obat yang berasal dari senyawa alam diklaim lebih manjur untuk pasien kanker dibandingkan dengan obat sintetik. Untuk saat ini, diketahui bahwa 10,000 dari 50,000 spesies tanaman yang mengandung khasiat obat, tersebar di berbagai ekosistem, namun hanya sebagian dari tanaman tersebut telah dianalisa dan diinvestigasi untuk potensi obat terapi. Di Malaysia, studi tanaman herbal sebagai sumber potensial dari kanker terapi sedang dikembangkan. Di antara tanaman-tanaman yang sedang diinvestigasi untuk agen terapi potensial adalah Keladi tikus (Typhonium flagelliforme) untuk pengobatan leukemia (Mohan et al., 2010(a); Mohan et al., 2010(b)), chalcone dari Temu Kunci (Boesenbergia rotunda) untuk kanker paru-paru (Isa et al., 2012), dan Tapak Liman (Elephantopus scaber) untuk kanker payudara (Ho et al., 2011). Beberapa zat aktif lainnya yang diinvestigasi adalah girinimbine dari akar dari Salam Koja (Murraya koenigii) untuk pengobatan kanker hati (Syam et al., 2011), dentatin dari daun Sicerek (Clausena excavate burm) untuk kanker prostat (Arbab et al., 2013), minyak biji dari Kenaf (Hibiscus cannabinus) dan phenylbutenoids dari Bangle (Zingiber cassumunar roxb) untuk leukemia (Foo et al., 2011; Anassamy et al., 2013). Indonesia setidaknya memiliki 9,600 dari sekitar 30,000 spesies yang mempunyai aktivitas farmakologis (Depkes RI, 2007). Penelitian sebelumnya telah membuat basis data senyawa herbal yang berasal dari Indonesia, Basis data tersebut menyediakan informasi dari struktur tiga dimensi dari senyawa herbal yang dapat diakses di tautan berikut: http://herbaldb.farmasi. ui.ac.id. Penelitian sebelumnya juga meneliti senyawa herbal yang terdapat di Vietnam dan mengumpulkan struktur kimia dari senyawa tersebut dari situs Pubchem dan Chemspider database (http://pubchem.ncbi.nlm.nih.gov dan http://www. chemspider. com/) (Ngo dan Li, 2013). Dalam mendesain suatu obat, tanaman-tanaman herbal ini menjadi bahan dasar untuk dipelajari lebih jauh. Penyaringan dari kegiatan farmakologi dari zat aktif tanaman obat memerlukan biaya yang besar, sumber daya manusia yang handal, dan waktu yang lama jika dilakukan tanpa arah yang jelas (Jayaram, 2011). Metode in silico atau ilmu bioinformatika dapat membuat simulasi dan perhitungan didalam desain obat. Metode yang digunakan adalah Computer Assisted Drug Design (CADD) yang mengkaji interaksi antara kandidat obat dan target secara komputasi (Hawkins dan Skillman, 2006). Molecular docking adalah metode umum yang digunakan. Molecular docking digunakan untuk memprediksi kompleks intermolekuler antara molekul obat dan target protein. Data yang dibutuhkan terdiri dari informasi ligan, atau obat yang akan dianalisis, dan reseptor atau protein target. Informasi yang dibutuhkan pada proses ini merupakan struktur tiga dimensi dari ligan dan reseptor (Abraham, 2003). Kemudian, ilmu bioinformatika juga menggunakan pendekatan machine learning untuk mempelajari kanker dan virus, sehingga hal ini dapat menjadi modal informasi yang baik untuk pengembangan obat berbasis herbal (Parikesit, 2018; Parikesit et al., 2018). Telaah sistematis ini akan membahas studi kasus yang memanfaatkan senyawa herbal untuk pengobatan kanker berbasis metode machine learning. Salah satu contoh yang akan didalami adalah Lysine-specific demethylase 1 (LSD1) yang telah diketahui mengalami ekspresi yang lebih tinggi di banyak jenis sel kanker, diantaranya pada kanker payudara (Lim et al., 2010), neuroblastoma (Schulte et al., 2009), dan kanker prostat (Wissman et al., 2007). Dengan demikian, penghambatan terhadap LSD1 oleh molekul lainnya adalah strategi yang efektif untuk menurunkan ekspresi gen yang terlibat pada proses diferensiasi, migrasi, dan invasi dari sel kanker atau untuk menghambat jalur proliferatif di berbagai kanker (Mohammad et al., 2013; Schmitt et al., 2013). Xanthones adalah senyawa polifenol alami dengan beragam aktivitas biologi, biokimia, dan farmakologi yang ditemukan di berbagai macam tanaman (Negi et al., 2013). Salah satu jenis dari xanthones, αmangostin adalah konstituen utama yang diisolasi dari buah manggis (Garcinia mangostana). Khasiat biologis yang signifikan menjadikannya sebagai kandidat bahan untuk pengembangan obat yang baru, khususnya untuk obat anti-kanker (Ibrahim et al., 2015). Dengan demikian, telaah ini menjadikan α-mangostin sebagai inhibitor terhadap LSD1. Kajian bioinformatika, diperkuat dengan kajian laboratorium basah, telah dilakukan dalam mengkaji beberapa senyawa bahan alam dan turunannya sebagai inhibitor LSD1. Diantaranya adalah sintesis turunan Xanthine (Ma et al., 2019); kajian simulasi molekular terhadap turunan senyawa 6-aril-5-sianopirimidin (Ding et al., 2017); dan simulasi molekuler terhadap relasi ROR α terhadap LSD1 (Kim et al., 2017). Keberhasilan penemuan ligan untuk inhibitor LSD1 dengan metode komputasi, terutama yang berbasis bahan alam dan turunannya, merupakan motivasi utama dalam kajian telaah sistematis ini. Sehingga diharapkan dapat memberikan narasi yang jelas mengenai detail teknis dari metode machine learning yang digunakan dalam simulasi tersebut dalam contoh interaksi reseptor-ligan dalam kasus lainnya secara lebih umum.

Navigasi Koleksi Struktur 3D dari Bahan Alam Langkah pertama untuk mendesain obat molekuler dari bahan alam adalah pengumpulan informasi terhadap bahan alam yang dapat digunakan sebagai obat. Informasi tersebut dapat ditemukan pada literatur, jurnal, buku dan situssitus. Seperti yang telah dijelaskan dalam telaah Parikesit et al. (2018), langkah ini dimulai dengan mengumpulkan informasi mengenai bahan kimia yang ditemukan di tanaman obat beserta struktur 2D dari berbagai literatur dan situs Pubchem. Kemudian, software PyMOL digunakan untuk menghasilkan struktur 3-dimensi dan menyimpannya kedalam basis data. Informasinya tersedia di http://herbaldb.farmasi.ui.ac.id. Lebih dari 1412 struktur 3-dimensi dari bahan kimia dari bahan alam di Indonesia yang disimpan di situs http://herbaldb.farmasi.ui.ac.id yang dapat diakses oleh publik dengan persetujuan admin untuk penelitian (Yanuar et al., 2011). Analisis in silico atau Bioinformatika Setelah mengetahui struktur dari bahan alam dan target protein, pengujian atau analisis in silico dapat diterapkan untuk melihat seberapa signifikan pengaruh bahan alam tersebut terhadap target protein. Dari penelitian yang dilakukan oleh Han et al. (2018), langkah-langkah yang dilakukan untuk mengetahui efek dari α-mangostin terhadap LSD1 dalam aktivitas pengobatan kanker adalah analisis penambatan molekul dan pendekatan machine learning.

Penambatan Molekul Analisis penambatan molekul dilakukan menggunakan software MOE 2009. Dalam kasus LSD1, Struktur dari LSD1 (PDB ID: 2V1D) dengan FAD (flavin adenine dinucleotide) bebas, sebuah peptida yang mirip dengan histone H3k4 methyltransferase sebagai substratnya, dan CoREST (Protein REST coprepressor 1) sebagai korepresor dipilih sebagai protein untuk analisis docking. Posisi grup dihidroksil dari rangka xanthone membentuk ikatan hidrogen dengan rantai samping dari Arg316 dan Thr810. Cincin benzen membentuk interaksi π-π dengan Tyr761, Glu801, Ala331, Val811, dan Arg316. Dua kelompok isopentene diprediksi berikatan dengan interaksi hidropobik dengan Leu659, Trp751, dan Ala814.

Dari analisis penambatan molekul diatas, rangka xanthone ditemukan secara signifikan berinteraksi dengan LSD1 dan kedua kelompok isopentene dapat berfungsi sebagai kelompok fungsional aktif untuk mempengaruhi bioaktivitas dari LSD1 (Han et al., 2018). Tidak hanya pada LSD1, metode ini juga dapat digunakan untuk mengkaji penambatan molekul ligan-protein lainnya.

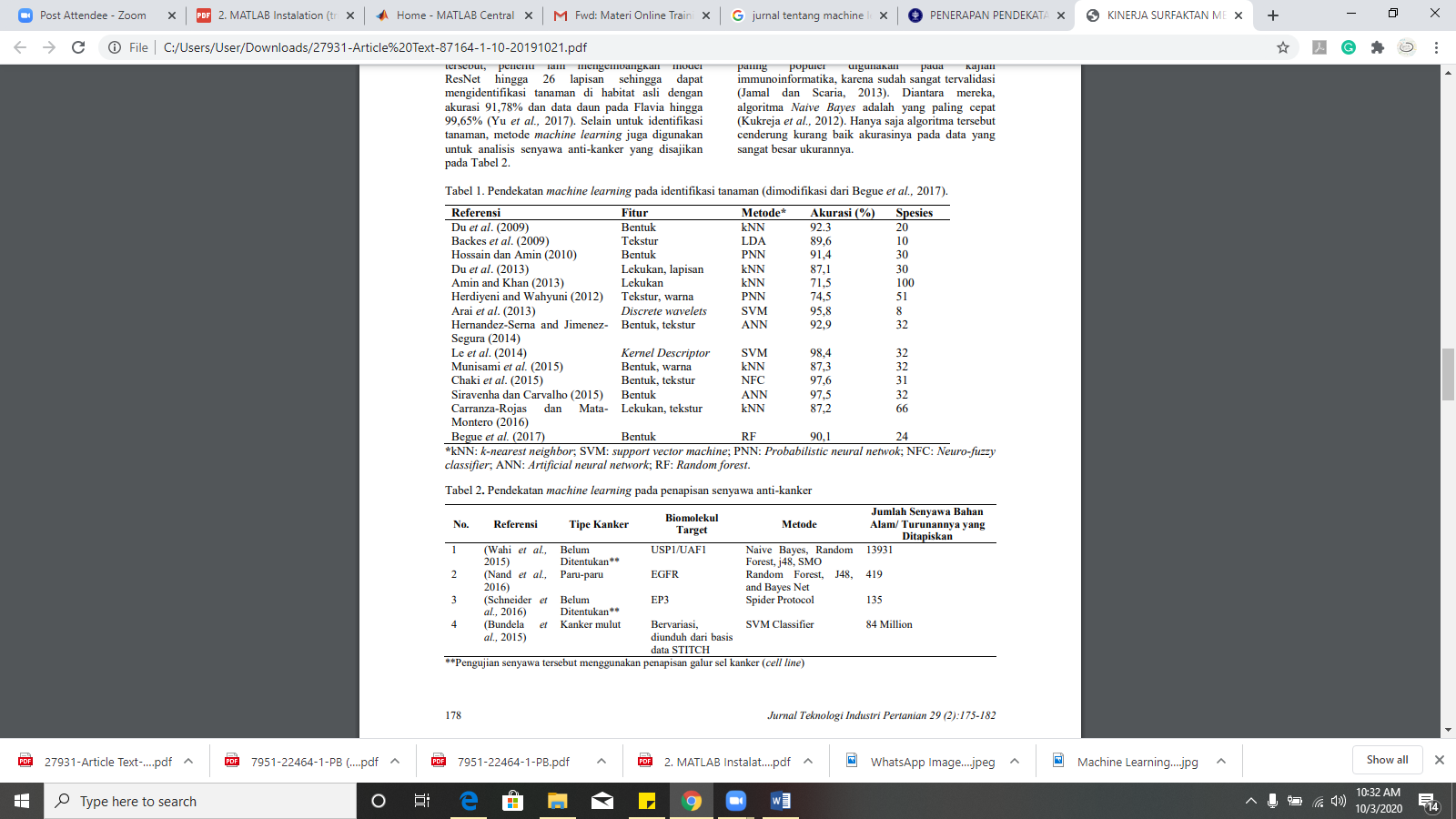
Pendekatan Machine Learning Penggunaan pendekatan machine learning telah digunakan pada berbagai aspek penelitian yang berhubungan dengan pengembangan obat. Machine learning adalah pendekatan prediktif yang mampu memprediksi luaran suatu sistem berbasis pada pola yang sudah ada. Runutan dalam metode machine learning dijabarkan sebagai berikut: Kumpulan besar data senyawa dari basis data bahan alam yang tersedia untuk umum diambil (proses data mining) untuk mengidentifikasi senyawa terapi potensial untuk kanker. Jutaan senyawa dapat disaring untuk mengkaji aktivitas anti-kanker oleh custom build classifier SVM. Target molekuler senyawa antikanker yang diprediksi didapatkan dari sumber yang dapat diandalkan, seperti penelitian bioassay eksperimental yang terkait dengan senyawa tersebut dan dari database interaksi protein-ligan. Basis data senyawa yang dapat digunakan sebagai terapi dan daftar senyawa anti-kanker bahan alam yang berasal dari studi literatur digunakan untuk membangun model regresi kuadrat parsial terkecil. Model regresi yang dibangun digunakan untuk estimasi bobot spesifik kanker berdasarkan target molekuler.

Bobot ini digunakan untuk menghitung skor untuk menyaring senyawa anti-kanker yang diprediksi memiliki potensi dalam mengobati kanker (Bundela et al., 2015). Pada telaah sistematis ini, beberapa penelitian yang sesuai telah ditemukan dalam kurun waktu 10 tahun terakhir. Beberapa dilaporkan dengan penggunaan data visual dan natural language processing dengan menerapkan konsep data mining untuk melakukan identifikasi tanaman obat maupun kandungan zat aktif yang dimiliki. Penggunaan Machine Learning Dalam Identifikasi Tanaman Obat Proses identifikasi dan klasifikasi spesies yang memiliki khasiat obat merupakan langkah awal dari proses penemuan obat. Proses identifikasi membutuhkan tenaga ahli yang memahami penanda yang dimiliki oleh spesies, baik secara morfologi maupun genetis. Sejauh ini, proses yang paling sukses dalam identifikasi adalah menggunakan pendekatan morfologi, meskipun melalui proses yang cukup panjang (Begue et al., 2017).

Pendekatan machine learning telah digunakan dalam melakukan proses identifikasi tanaman berikut ini (Tabel 1). Dalam kurun 10 tahun, pendekatan machine learning telah membantu dalam proses identifikasi secara visual dan bahkan telah berhasil diterapkan untuk mendeteksi lesi, mitosis, dan status gen pada kanker payudara (Ramanto dan Parikesit, 2019). Sehingga validitas dan reliabilitasnya tak perlu diragukan lagi. Berdasarkan tabel 1, terdapat metode machine learning yang berbeda dan memberikan akurasi yang cukup baik dalam mengidentifikasi spesies yang digunakan dalam proses training. Du et al. (2009) menggunakan metode kNN untuk mengidentifikasi 2000 foto dari 20 spesies dan mendapatkan 92,3% akurasi. Penelitian lainnya menggunakan metode seperti PNN, ANN, dan SVM yang memberikan akurasi melebihi 90% (Hossain dan Amin, 2010; Le et al., 2014; Chaki et al., 2015).

Salah satu metode yang dikembangkan bahkan dibuat menjadi aplikasi untuk smartphone seperti MedLeaf (Prasvita dan Herdiyeni, 2013). Metode machine learning dapat diterapkan pada tanaman yang berasal dari manapun, seperti Indonesia (Herdiyeni dan Wahyuni, 2012; Prasvita dan Herdiyeni, 2013), Vietnam (Le et al., 2014), bahkan tanaman ornamental (Arai et al., 2013) selama basis data yang digunakan untuk “melatih” program memiliki informasi yang cukup. Uniknya, kebanyakan dari program yang dibuat, menggunakan data yang sama dengan data yang digunakan untuk “melatih” program Flavia yang berisi 33 gambar daun tanaman yang diketahui memiliki khasiat obat (Wu et al., 2007). Dengan menggunakan dataset tersebut, peneliti lain mengembangkan model ResNet hingga 26 lapisan sehingga dapat mengidentifikasi tanaman di habitat asli dengan akurasi 91,78% dan data daun pada Flavia hingga 99,65% (Yu et al., 2017).

Selain untuk identifikasi tanaman, metode machine learning juga digunakan untuk analisis senyawa anti-kanker yang disajikan pada Tabel 2. Berdasarkan Tabel 2, metode berbasis machine learning pada dasarnya terbagi dua, yaitu untuk data senyawa berukuran kecil/sedang (nomor 1-3), dan berukuran besar (nomor 4). Salah satu alasan mengapa SVM classfier merupakan metode yang sangat cocok untuk data berukuran sangat besar adalah karena metode tersebut secara ketat membangun model berdasarkan contoh dari data pelatihan. Selanjutnya, data tersebut digunakan untuk memprediksi nilai target dari contoh data uji dan hanya diberikan atribut dalam data uji (Bundela et al., 2015). Sementara itu, metode SPIDER menggunakan self organizing map (SOM) yang berbasis algoritma klaster sehingga cenderung membutuhkan daya komputasi besar untuk set data yang kecil (Reker et al., 2014). Naive Bayes, Random Forest, j48, SMO adalah algoritma yang paling populer digunakan pada kajian immunoinformatika, karena sudah sangat tervalidasi (Jamal dan Scaria, 2013). Diantara mereka, algoritma Naive Bayes adalah yang paling cepat (Kukreja et al., 2012). Hanya saja algoritma tersebut cenderung kurang baik akurasinya pada data yang sangat besar ukurannya.



Penggunaan Machine Learning Meningkatkan Efisiensi Penemuan Obat Penggunaan model prediktif dengan menggunakan komputer dapat membantu memprediksi hasil eksperimen yang bahkan belum dilaksanakan. Bila model tersebut telah dibuat, maka perlu adanya eksperimen lain untuk mengetahui akurasi dari model. Pembuatan model tersebut harus secara teliti karena hasil false-positive maupun falsenegative dapat memberikan kerugian yang signifikan. (Murphy et al., 2013). Beberapa pendekatan dilakukan untuk meningkatkan akurasi dari model prediktif. Salah satunya adalah algoritma active learning yang mengedepankan otomatisasi dalam proses “pelatihan”. Metode ini dapat memprediksi 60% adanya reaksi dari data senyawa dan protein setelah melakukan eksplorasi sebanyak 3% dari total 3,54 juta kemungkinan, jauh lebih baik dari model greedy selection with QSAR-like model dan Computational Research Engine. Uniknya, model ini akan semakin baik seiring dengan semakin banyaknya data yang diberikan (Naik et al., 2013). Model machine learning juga dibuat untuk mengidentifikasi adanya ikatan antara senyawa aktif dengan protein. Berbeda dengan sebelumnya, model ini menggunakan cascaded learning method yang terdiri atas 2 tahapan. Tahapan pertama menggunakan metode yang menyeleksi binding activity dari senyawa berdasarkan strukturnya. Hal ini didasarkan pada metode structure-activityrelatonship (SAR) yang dapat digunakan untuk memprediksi ikatan dari suatu senyawa, selama diketahui strukturnya. Tahapan kedua menggunakan data SAR dari tahap pertama untuk melakukan seleksi selektifitas dari senyawa yang diuji. Tahapan kedua ini akan meningkatkan akurasi dengan menggunakan metode machine learning. Model ini dibuat dengan metode Neural network dengan penambahan multi-tasking, cascaded, dan three way yang dibandingkan dengan baseline model structure-selectivity-relationship (SSR). Penambahan multi-tasking, cascaded, dan three way meningkatkan performa dari model baseline, terutama pada skema target protein tunggal. Meskipun begitu, pada skema target lebih dari satu, performanya menurun sehingga diperlukan kajian lebih lanjut (Ning dan Karypis, 2012).

Salah satu tahapan penting dalam kajian pengembangan obat adalah mulai digunakannya metode deep learning. Sebagai bagian dari metode machine learning, metode ini menggunakan basis artficial neural network yang terinspirasi dari simulasi jaringan saraf di otak manusia maupun bagian tubuh lainnya (Schmidhuber, 2015). Aplikasi deep learning dalam penapisan senyawa anti-kanker memang belum sebanyak dan sebesar metode machine learning lainnya karena kajian ini masih sangat baru, namun sudah mulai ada pengembangan dan penggunaanya. Terdapat beberapa aplikasi deep learning di kajian ini. Penggunaan profil obat kanker sebagai model deep learning yang memprediksi efektifitas obat kanker secara genomik merupakan salah satu aplikasi pada ranah personalized medicine dengan menggunakan data galur sel manusia dan struktur obat untuk melakukan simulasi molekuler (Chang et al., 2018).

Kemudian, penggunaan deep learning untuk mengetahui efek sinergistik dari obat kanker dengan menggunakan data chemo dan genomics informatics (Preuer et al., 2018). Perbedaan aplikasi deep learning dengan metode machine learning lainnya seperti Random Forest dan SVM adalah penggunaan indikator data yang lebih banyak dan lebih kompleks dalam konteks jaringan (networking). Deep learning lebih cocok diaplikasikan untuk struktur data yang kompleks dan terkoneksi satu sama lain. Hal ini terbukti dengan efektifitas penggunaannya pada crowd-sourced QSAR dan prediksi toksikologi terhadap senyawa bioaktif, yang melibatkan banyak nodus di jaringan internet (Koutsoukas et al., 2017).

KESIMPULAN DAN SARAN Kesimpulan Metode machine learning ternyata sangat bermanfaat untuk melakukan penjaringan kepustakaan senyawa bahan alam dan turunannya, baik data berukuran kecil/sedang, maupun berukuran besar (big data). Diketahui bahwa metode machine learning paling efektif dan efisien adalah Naïve Bayes. Namun yang terbaik dalam mengolah data senyawa dalam jumlah besar adalah SVM classfier.

Metode deep learning sudah mulai digunakan untuk komputasi data senyawa bioaktif yang kompleks, walaupun aplikasinya masih terbatas. Saran Pengembangan basis data bahan alam dapat diarahkan untuk menjadi repositori data wet laboratory bahan alam dengan kapabilitasi prediktif berbasis machine learning, terutama deep learning sehingga bisa menjadi pusat rujukan anotasi untuk eksperimen berikutnya. Jika data yang ada strukturnya sangat kompleks dan membentuk suatu topologi jaringan, maka disarankan menggunakan metode deep learning untuk komputasinya. Untuk memperdalam telaah dalam kajian deep learning, dapat dilakukan penerapan metode Prisma untuk kedepannya.

Quiz Pertemuan 5

Berikan contoh penerapan algorima yang ada di machine learning dengan penyelesaian studi kasusnya, selain yang sudah di contohkan pada modul.

# Daftar Pustaka

[1] “Data Scientist: The Sexiest Job of The 21st Century” Diakses 01 April 2019 dari https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

[2] “50 Best Jobs in America 2019” Diakses 01 April 2019 dari https://www.glassdoor.com/List/Best-Jobs-in-America-LST\_KQ0,20.htm

[3] “Tech in Asia Jobs” Diakses 02 Desember 2018 dari https://www.techinasia.com/jobs

[4] “Kalibrr: Where Jobs Find You” Diakses 02 Desember 2018 dari https://www.kalibrr.com/

[5] “Data Scientist Earning More Than CAs, Engineers” Diakses 01 April 2019 dari https://timesofindia.indiatimes.com/india/Data-scientists-earning-more-than-CAs-engineers/articleshow/52171064.cms

[6] “5 Facts About Software Engineers, Like Which One Gets Paid The Most” Diakses 01 Apil 2019 dari https://learning.linkedin.com/blog/tech-tips/the-american-city-that-pays-software-engineers-the-most–and-oth

[7] “Top 10 Reasons Why You Should Learn Data Analytics” Diakses 01 April 2019 dari https://bigdata-madesimple.com/10-reasons-why-you-should-learn-data-analytics/

[8] “5 Reasons Why Everybody Should Learn Data Analytics” Diakses 01 April 2019 dari https://www.sas.com/en\_au/insights/articles/analytics/5-reasons-why-everybody-should-learn-data-analytics.html

[9] “Permintaan Tenaga Data Scientist Melonjak, Jadikan Profesi Ini Kian Menjanjikan” Diakses 10 Desember 2018 dari https://id.techinasia.com/talk/profesi-data-scientist-menjanjikan

[10] “Go-Jek Buka Kantor Data Science di Singapura, Apa Alasannya?” Diakses 01 April 2019 dari https://www.liputan6.com/tekno/read/2998722/go-jek-buka-kantor-data-science-di-singapura-apa-alasannya

[11] Introduction to Data Science. A Python Approach to Concepts,Techniques and Applications. Laura Igual, Santi Segui. Tahun

2017.

[12] Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk from The Frontline. O'Reilly. 2014.

[13] Jure Leskovek, Anand Rajaraman and Jerey Ullman. Mining of Massive Datasets. v2.1, Cambridge University Press. 2014. (free online)

[14] Mohammed J. Zaki andWagner Miera Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press. 2014.

[15] https://anaktik.com/data-science/

[16] https://anaktik.com/skill-data-scientist/