2. LANDASAN TEORI

2.1. Tinjauan Pustaka

2.1.1. Diagnosis

"Diagnosa adalah istilah yang merujuk pada pemeriksaan terhadap suatu hal" (Abdi, November 29, 2021). Dalam penelitian ini, diagnosis memiliki arti hasil pemeriksaan terhadap seseorang berdasarkan gejala yang ditemukan pada saat pemeriksaan. Dalam tindakan medis, pemberian diagnosis terhadap seorang pasien dilakukan oleh dokter yang berpengalaman dan memiliki kemampuan untuk menentukan penyakit yang diderita oleh pasien.

Ketika seorang dokter memberikan sebuah diagnosis yang kurang tepat, maka dapat menimbulkan misdiagnosis, dimana pemberian keputusan penyakit memiliki kekurangan (*underdiagnosis*) atau kelebihan (*overdiagnosis*), (W. Rita, personal communication, April 28, 2023). Sebuah misdiagnosis sendiri merupakan tindakan berbahaya yang sangat penting karena dapat mengakibatkan ketidaksembuhan pasien maupun tindakan hukum terhadap dokter yang memberinya. Oleh karena itu, perlu adanya cara yang dapat membantu dalam proses verifikasi dalam pemberian sebuah diagnosis.

2.1.2. Gejala

Menurut Kamus Besar Bahasa Indonesia (2023), "Gejala" diartikan sebagai sebuah perihal yang tidak biasa dan perlu diperhatikan dimana perihal ini dapat menandakan timbulnya sesuatu. Dalam penelitian skripsi ini, gejala memiliki peranan penting sebagai objek penelitian dikarenakan diagnosa dan obat yang diberikan akan bergantung pada gejala yang dimiliki oleh pasien.

2.1.3. Naïve Bayes

Metode Naive Bayes adalah sebuah set dari algoritma *Supervised Learning* dengan menerapkan teorema Bayes dengan asumsi "*naive*" untuk menghitung probabilitas kondisional (scikit-learn.org, nd). Proses kalkulasi dilakukan dengan menggunakan probabilitas dari terjadinya sesuatu berdasarkan kondisi yang dapat mempengaruhi *outcome* tersebut (Sridhar et al., 2023). Berdasarkan penelitian yang dilakukan oleh Sridhar et al., (2023), penggunaan metode Naive Bayes menghasilkan nilai akurasi terendah dibandingkan model lainnya. Namun hasil penelitian ini

berbeda dengan penelitian yang dilakukan oleh Mahoto et al., (2022), dimana model Naive Bayes berhasil mencapai nilai akurasi terbaik lebih banyak dibandingkan model lainnya. Berdasarkan perbedaan kedua hasil penelitian ini, penelitian skripsi diharapkan dapat menentukan model terbaik dengan menggunakan data yang berbeda dari data yang digunakan oleh kedua penelitian sebelumnya. Pengambilan *class label* dilakukan dengan mengambil kelas dengan probabilitas tertinggi dengan menerapkan teorema Bayes (S. Ray, February 6, 2024).

$$P(c|x) = \frac{P(x|c) \times P(c)}{P(x)}$$

Dengan :

- Probabilitas Posterior (P(C|X)): Probabilitas bahwa suatu contoh termasuk dalam kelas tertentu C setelah melihat fitur X
- Kelas (C): Kelas yang diprediksi untuk suatu contoh setelah melihat fitur X
- Fitur (X): Data yang diamati atau fitur yang digunakan untuk memprediksi kelas
- Probabilitas Prior (P(C)): Probabilitas murni dari kelas C sebelum melihat data
- Probabilitas Kondisional (P(X|C)): Probabilitas munculnya fitur X dalam kelas C

2.1.4. Decision Tree

Decision Tree merupakan metode *supervised machine learning* yang digunakan untuk klasifikasi dan regresi, dengan tujuan menghasilkan model yang dapat memprediksi nilai dari *target variable* dengan mempelajari aturan pengambilan keputusan yang didapat dari *data features* (scikit-learn.org, nd). Hasil akhir dari algoritma Decision Tree sangat bergantung terhadap *node* awal yang ditetapkan sebagai *root* yang dapat dipilih menggunakan penghitungan *Information Gain* (*IG*) dan *Gini Index* (*GI*), (Mahoto et al., 2022). Hasil yang didapat dari penggunaan Decision Tree sendiri tidak terlalu berbeda bila dibandingkan dengan metode yang lain (Sridhar et al., 2023), dan tidak mencetak nilai terbaik dengan jumlah banyak (Mahoto et al., 2022). Meskipun demikian, berdasarkan penelitian yang dilakukan Shehab et al., (2022), Decision Tree yang digunakan dalam sebuah *ensemble method*

dapat menghasilkan hasil yang lebih akurat dibandingkan bila metode tersebut hanya berjalan sendiri.

Decision tree bekerja dengan menggunakan *entropy* atau penghitungan dari *impurity*, dimana akan didapatkan *information gain* dari pembandingan sebuah atribut sebelum dan sesudah *splitting* untuk menentukan *parent node* dari *tree* yang akan dibentuk (H. Juwiantho, August 29, 2022).

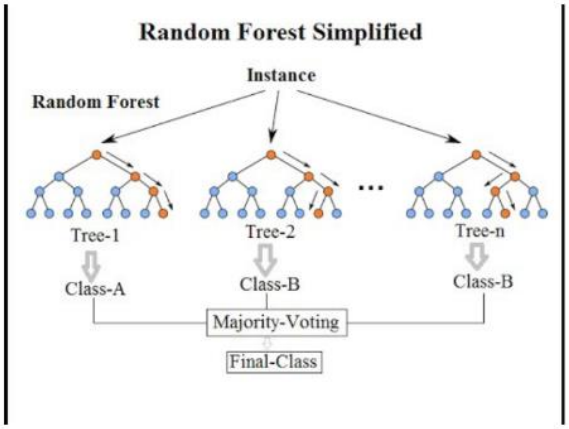
$$Entropy(t) = -\sum_{j} p(j|t) \log p(j|t)$$

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} Entropy(i) \right)$$

2.1.5. Random Forest

Random Forest merupakan sebuah *meta estimator* yang menerapkan beberapa Decision Tree terhadap berbagai sub-sampel dari *dataset*, serta menggunakan rata-rata untuk meningkatkan akurasi prediksi dan mengontrol *overfitting* (scikit-learn.org, nd). Dalam melakukan proses prediksi, Random Forest sendiri telah memiliki akurasi diatas batas pengujian (90%) dimana akurasi dapat mencapai 93% dalam train data dan 95% dalam test data (Sridhar et al., 2023). Dalam riset yang dilakukan oleh Shehab et al., (2022), penulis menjelaskan bahwa dalam pengujian tiga *dataset*, algoritma Random Forest sangat membantu dalam membuat sebuah sistem diagnosa yang dibantu komputer

algoritma Random Forest merupakan ekstensi dari penggunaan *ensemble method*, Bagging dengan membuat sebuah "*forest*" dari beberapa Decision Tree dengan menerapkan *feature randomness* (ibm.com, nd). Dari hasil prediksi *class* yang dilakukan oleh masing-masing *tree*, kemudian akan melakukan *majority voting* untuk menghasilkan sebuah *final class* sebagai *output class* (G. S. Budhi, April 18, 2022).



Gambar 2.1. Cara Kerja Algoritma Random Forest

2.1.6. Support Vector Machine

SVM merupakan sebuah algoritma *supervised machine learning* yang dapat digunakan dalam klasifikasi maupun regresi, dengan tujuan memberikan sebuah garis sebagai pembatas antara keputusan agar data baru yang dimasukan akan lebih gampang dikategorikan (javatpoint.com, nd). Berdasarkan riset yang dilakukan oleh Mahoto et al., (2022), metode SVM juga mampu memiliki tingkat performa yang lebih baik dibandingkan algoritma lainnya. Namun, berdasarkan hasil penelitian yang dilakukan oleh Sridhar et al., (2023), metode Svm memiliki akurasi yang sama dengan metode lainya dan lebih rendah dibandingkan model yang diusulkan.

2.1.7. Extreme Gradient Boosting

XGBoost merupakan salah satu sub-class dari Supervised machine learning yang dapat digunakan untuk klasifikasi dan regresi yang menggunakan prinsip Decision Tree dimana *node* splitting ditentukan oleh *similarity score* dan *gain*, (towardsdatascience.com, February 5, 2022).

2.1.8. Evaluation Matrix

Evaluation matrix merupakan matriks yang terdiri dari *accuracy, precision*, dan *recall*.

2.1.8.1. *Accuracy* merupakan salah satu cara menghitung performa model klasifikasi, yang dilakukan dengan membagi total prediksi yang benar dengan total prediksi (Google, n.d.). *accuracy* dapat dinilai berdasarkan total dari prediksi yang tepat dibagi dengan semua prediksi (evidentlyai.com, n.d.).

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Predictions}$$

2.1.8.2 *Precision* merupakan cara penilaian performa algoritma yang bertujuan untuk menunjukan proporsi dari proses identifikasi yang benar, dengan cara menghitung *True Positive* yang dibagi dengan *True Positive* + *False Positive* (Google, n.d.). dengan pengertian lain, *precision* digunakan untuk menghitung seberapa banyak prediksi positif yang benar (evidentlyai.com, n.d.).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

2.1.8.3. *Recall* bertujuan untuk menentukan proporsi dari nilai positif yang berhasil diidentifikasi sebagai positif dengan penghitungan *True Positive* yang dibagi oleh *True Positive* + *False Negative* (Google, n.d.). dengan pengertian lain, *recall* digunakan untuk mengetahui apakah sebuah model dapat menumakan semua kasus dari *class* tertentu (evidentlyai.com, n.d.).

$$Recall = \frac{True\ Positive}{True\ Positive + false\ Negative}$$

2.1.9. Ensemble Method

Ensemble method adalah kumpulan teknik untuk membuat gabungan model untuk menghasilkan hasil yang lebih baik, dimana penggunaan *ensemble method* dalam *machine learning* dapat menghasilkan hasil yang lebih akurat daripada jika hanya menggunakan sebuah model (Demir, February 4, 2016). Penerapan *ensemble method* dalam skripsi ini digunakan dalam metode Bagging yang didasarkan pada penelitian terkait, dimana metode Bagging dapat menghasilkan nilai akurasi yang lebih baik dan bervariasi dibandingkan dengan metode voting dan model dasar (Mahoto et al., 2022).

2.1.10. Voting

Ensemble Voting merupakan sebuah teknik *ensemble method* yang digunakan dengan menggabungkan *predictors* dari beberapa model individu (soulpageit.com, june 19, 2023). *Ensemble* Voting dibagi kedalam *majority* Voting, dimana sebuah label yang menjadi hasil mayoritas diambil menjadi *class label*, *weighted* Voting dimana setiap model mendapat nilai tertentu dari hasil prediksinya dan *class label* didapat dari menjumlahkan atau mengambil nilai rata-rata dari hasil prediksi, dan *soft* Voting dimana hasil akhir diambil dari class label dengan nilai rata-rata

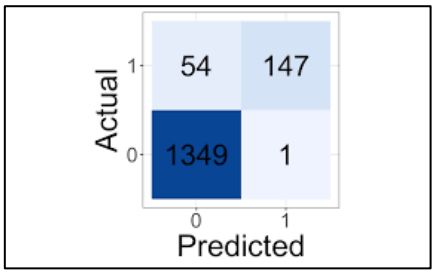
probabilitas tertinggi (soulpageit.com, june 19, 2023). Penerapan dari metode voting dilakukan menggunakan *function* VotingClassifier dari Library SK-Learn.

2.1.11. Bagging

Bagging (Bootstrap Aggregating) merupakan sebuah teknik *ensemble learning* yang digunakan untuk meningkatkan akurasi dan mengatasi masalah *bias-variance* dengan mengurangi *variance* dari sebuah model (Biswal, February 14, 2023). Bagging sendiri dapat dilakukan dengan dilakukan terhadap masing-masing model (Mahoto et al., 2022). Metode bagging dilakukan dengan menggunakan *function* BaggingClassifier dari Library SK-Learn.

2.1.12. Confusion matrix

Confusion matrix merupakan sebuah matriks berukuran N*N yang digunakan untuk menilai performa dari sebuah *classification model*, dengan N merupakan jumlah target classes (Suresh, A., June 22, 2021). “Seperti yang telah dijelaskan diatas, *confusion matrix* akan memberi tahu seberapa baik model yang kita buat. Secara khusus *confusion matrix* juga memberikan informasi tentang TP, FP, TN, dan FN.” (Nugroho, K. S., June 4, 2020). Dikarenakan peranan pentingnya sebagai sebuah metode untuk mengevaluasi model, maka penelitian skripsi ini menerapkan penggunaan *confusion matrix* sebagai salah satu tolak ukur penilaian dari model.



Gambar 2.2. Contoh dari *Confusion Matrix*

2.1.13. Dataset

Dataset merupakan sebuah *file* yang mengandung satu atau lebih informasi dasar yang dapat digunakan oleh sebuah program dalam sebuah *operating system* (ibm.com, n.d.). dalam penelitian skripsi, data yang digunakan berasal dari dua *dataset* (aplikasi rekam medis klinik dan website BPJS) yang kemudian di-integrasikan kedalam sebuah *dataset* dalam bentuk *file* .CSV yang berisikan tabel mengenai karakteristik pasien, gejala yang dialami pasien, diagnosa yang diberikan, dan juga obat yang diberikan.

2.1.14. Preprocessing

Preprocessing merupakan tahap pemberishan, transformasi, dan integrasi data untuk memperisapkan data agar dapat dianalisis (GfG, May 6, 2023). Pembersihan data merupakan proses memastikan akurasi data dengan mendeteksi kesalahan data dan memperbaiki atau menghapus data sesuai dengan debutuhan (Xeratic, August 3, 2022). Transformasi meliputi koversi dan mengatur struktur dari data kedalam format yang dapat digunakan dalam proses analisis (Tibco.com, n.d.). integrasi merupakan proses penggabungan data yang berasal dari dua atau lebih *database* yang berbeda

2.2. **Tinjauan Studi**

A. *Machine learning in medical applications: A review of state-of-the-art methods* (Shehab et al., 2022)

- Masalah yang diangkat dari penelitian ini adalah bagaimana algoritma *Machine Learning* dapat meningkatkan keandalan, performa, kemampuan memprediksi, dan akurasi dari sistem diagnosis terhadap beberapa penyakit. Penelitian ini ditujukan untuk menghasilkan sebuah tabel untuk menunjukan hasil atau performa dari algoritma terbaik dalam mendiagnosis sebuah kasus.
- Metode yang digunakan dalam penelitian ini meliputi Support Vector Machine (SVM), Naive Bayes (NB), Artificial Neural Network (ANN), k-Nearest Neighbor (K-NN), Decision Tree (DT), Back-Propagation Neural Network (BPNN), Support Vector Regression (SVR), Multiple Linear Regression (MLR), Partial Least Squares (PLS), k-Means, Hierarchical Algorithm (HA), Mean-Shift, Density-Based Spatial Clustering of Application with Noise (DBSCAN), Feature Selection, Feature Extraction, Q-learning, Temporal Difference, Value iteration, dan Markov Decision.
- Hasil dari penelitian ini adalah hasil penerapan *Machine Learning* terhadap 5 kasus yang disediakan, yakni; *Cancer*, *Medical chemistry*, *Brain*, *Medical Imaging*, dan *Wearable sensor*. Pada deteksi kanker, algoritma didominasi penggunaan *supervised machine learning* dalam rangka membantu klasifikasi dari input. Dalam deteksi gangguan dalam jaringan otak, algoritma *machine learning* dapat digabung dengan teknik *AI* untuk membantu melakukan tracking terhadap bagian-bagian penting. Dalam meningkatkan diversitas dari component

classifiers, peneliti menyarankan penerapan Fuzzy *classifier* untuk membantu menentukan diagnosis. Peneliti juga menyarankan penggunaan algoritma *unsupervised machine learning* dalam memproses data berupa gambar. Dan terakhir, sebuah metode dibutuhkan untuk memproses data yang didapatkan dari *wearable sensors*.

- Perbedaan dari penelitian yang dilakukan dan penelitian skripsi adalah skripsi ini hanya mengambil algoritma-algoritma *Supervised Machine Learning*, dimana skripsi ini bertujuan untuk melakukan klasifikasi terhadap diagnosis berdasarkan gejala-gejala yang tersedia dalam dataset. Tujuan dari skripsi juga lebih terfokus pada klasifikasi dimana penelitian yang dilakukan memiliki tujuan yang berbeda, tergantung pada 5 kasus yang diambil. Oleh karena itu, model yang digunakan merupakan algoritma-algoritma *Supervised Machine Learning*. Penelitian skripsi juga menggunakan dataset yang berbeda.

B. *A machine learning based data modeling for medical diagnosis* (Mahoto et al., 2022)

- Masalah yang diangkat dari penelitian ini adalah bagaimana melakukan sebuah proses optimisasi terhadap model *machine learning* yang digunakan untuk mengatasi data medis yang bersifat multidimensional. Penelitian ini ditujukan untuk menghasilkan tabel yang dapat menunjukan hasil terbaik yang didapat dari semua model yang dibuat, serta optimisasi yang dilakukan menggunakan metode ensemble. Peneliti juga melakukan proses kalkulasi hasil yang dibagi kedalam perbandingan jumlah *training* dan *validation data*.
- Metode yang digunakan dalam penelitian ini meliputi Decision Tree (DT), Naive Bayesian (NB), Multilayer Perceptron (MLP), Random Forest (RF), dan Support Vector Machine (SVM). penelitian ini juga menerapkan pembagian jumlah *training* dan *validation data* untuk membandingkan akurasi dari semua model yang digunakan. Penerapan ensemble method dalam bentuk voting dan bagging untuk membandingkan performa terhadap algoritma dasar. Data yang digunakan dalam penelitian ini merupakan dua buah dataset yang memiliki atribut kondisi/gejala serta identitas pasien mengenai gangguan jantung dan infeksi mata.
- Hasil penelitian ini berupa tabel yang menunjukan performa dari semua model yang dinilai berdasarkan *accuracy*, *precision*, *recall*, *F-measure*, dan ROC. berdasarkan hasil yang didapat, performa terbaik kebanyakan terdapat pada

pembagian data kedalam 80% training dan 20% validation. Dari segi metode ensemble, metode *voting* memiliki hasil yang tidak terlalu berbeda dibandingkan dengan algoritma awal yang digunakan, namun metode Bagging seringkali menghasilkan performa yang berbeda dibandingkan dengan algoritma awal dan memiliki hasil yang lebih bagus. Hasil penelitian juga menyimpulkan bahwa algoritma *machine learning* dapat melakukan proses klasifikasi lebih baik dari algoritma *deep learning*.

- Perbedaan dari penelitian yang dilakukan dan penelitian skripsi adalah perbedaan dari *dataset* yang digunakan, dimana *dataset* dari skripsi memiliki kemiripan dengan *dataset* kedua dari penelitian namun memiliki limitasi yang berbeda (dataset penelitian terfokus pada penyakit mata, dataset skripsi bersifat lebih general dengan penyakit umum) dan dari sumber yang berbeda. Skripsi juga akan mengambil algoritma *machine learning* sebagai fokus riset, dikarenakan penelitian juga telah membuktikan bahwa penggunaan algoritma *machine learning* lebih superior dibandingkan deep learning. Berdasarkan hasil yang didapat dari *ensemble method*, skripsi juga akan menerapkan penggunaan *ensemble method* Bagging dalam melakukan proses klasifikasi.

C. *Mobile Application Development for Disease Diagnosis based on Symptoms using Machine Learning Techniques* (Sridhar et al., 2023)

- Permasalahan yang diangkat dari penelitian ini adalah kurangnya tenaga kerja dalam bidang kesehatan yang mengakibatkan rendahnya akses dari masyarakat terhadap fasilitas kesehatan. Oleh karena itu, dibutuhkan sebuah aplikasi *mobile* yang dapat digunakan sebagai pengganti tenaga kerja dengan menggunakan algoritma *machine learning* dalam menentukan diagnosis kepada pasien sebagai usaha meringankan beban kurangnya tenaga kerja.
- Metode yang digunakan meliputi K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), dan Artificial Neural Network (ANN). metode kemudian diterapkan terhadap dataset yang diambil dari Kaggle yang berisikan gejala yang dialami serta hasil diagnosisnya. Penelitian ini juga menggunakan metode bagging yang dilakukan terhadap hasil keseluruhan model yang digunakan sebagai pembanding terhadap algoritma dasar.

- Hasil dari penelitian ini merupakan sebuah aplikasi yang berbasis kepada hasil dari algoritma yang diajukan oleh tim peneliti yang memiliki nilai performa 98.84% dalam akurasi. Aplikasi ini kemudian dapat digunakan oleh dokter dalam melakukan verifikasi terhadap diagnosis yang diberikan guna mempersingkat waktu pemeriksaan pasien.
- Perbedaan dari penelitian ini dan skripsi yang dilakukan adalah perbedaan dataset yang digunakan. Penelitian skripsi mengambil dataset yang didapat langsung dari apotek X, sementara penelitian ini mengambil dataset yang didapat dari Kaggle. Isi dari dataset sendiri juga memiliki perbedaan, dimana dataset Kaggle hanya memiliki isi gejala dan diagnosis sebagai kolom, sementara dataset penelitian juga melihat atribut lain seperti umur dan jenis kelamin dikarenakan efek yang diberikan oleh atribut-atribut tersebut.

Tabel 2.1.

Tabel Perbandingan State-Of-The-Art dari Penelitian Sebelumnya

No.	Judul	Penulis	Tahun	Hasil	Kelebihan
1	*Machine learning in medical applications: A review of state-of-the-art methods*	Mohammad Shehab, Laith Abualigah, Qusai Shambour, Muhannad A. Abu-Hashem, Mohd Khaled Yousef Shambour, Ahmed Izzat Alsalibi, Amir H. Gandomi	2022	• Deteksi kanker, disarankan menggunakan *supervised machine learning* • Deteksi gangguan jaringan otak : menggunakan tambahan teknik *AI* untuk membantu penekanan	• Tujuan dari skripsi juga lebih terfokus pada klasifikasi. • model yang digunakan merupakan algoritma-algoritma *Supervised Machine Learning* • Penelitian skripsi juga menggunakan

				dalam bagian penting • Meningkatkan diversitas : disarankan penambahan *fuzzy classifier* • *Unsuperfised machine learning* sebagai pembantu dalam memproses gambar • Dibutuhkan metode bagi *wearable sensors*	dataset yang berbeda.
2	*A machine learning based data modeling for medical diagnosis*	Naeem Ahmed Mahoto, Asadullah Shaikh, Adel Sulaiman, Mana Saleh Al Reshan, Adel Rajab, Khairan Rajab	2022	• Hasil penelitian ini berupa tabel yang menunjukan performa dari semua model yang dinilai berdasarkan accuracy, precision, recall, F-measure, dan ROC.	• dataset penelitian terfokus pada penyakit mata, dataset skripsi bersifat lebih general dengan penyakit umum • Skripsi juga akan mengambil algoritma

				• voting memiliki hasil yang tidak terlalu berbeda dibandingkan dengan algoritma awal yang digunakan • metode Bagging seringkali menghasilkan performa yang berbeda dibandingkan dengan algoritma awal dan memiliki hasil yang lebih bagus • Hasil penelitian juga menyimpulkan bahwa algoritma machine learning dapat melakukan proses klasifikasi lebih baik dari	machine learning sebagai fokus riset, dikarenakan penelitian juga telah membuktikan bahwa penggunaan algoritma machine learning lebih superior dibandingkan deep learning

				algoritma deep learning.	
3	*Mobile Application Development for Disease Diagnosis based on Symptoms using Machine Learning Techniques*	Anirudh Sridhara, Ahmed Mawiaa, A. L. Amuthaa	2022	• Hasil dari penelitian ini merupakan sebuah aplikasi yang berbasis kepada hasil dari algoritma yang diajukan oleh tim peneliti yang memiliki nilai performa 98.84% dalam akurasi.	• Penelitian skripsi mengambil dataset yang didapat langsung dari hasil pemeriksaan Dr. Rita Wey • dataset penelitian juga melihat atribut lain seperti umur dan jenis kelamin dikarenakan efek yang diberikan oleh atribut-atribut tersebut