BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Untuk memperkuat dasar teoritis dan empiris penelitian ini, dilakukan kajian terhadap beberapa penelitian terdahulu yang relevan dengan topik prediksi depresi dan penerapan *machine learning*. Penelitian-penelitian tersebut membahas berbagai pendekatan dalam pemodelan klasifikasi, termasuk penggunaan *Random Forest*, teknik *ensemble*, serta metode optimasi dalam meningkatkan performa model prediksi pada bidang kesehatan mental maupun bidang lainnya. Tinjauan ini bertujuan untuk memahami metode yang telah digunakan, membandingkan hasil penelitian sebelumnya, serta mengidentifikasi celah penelitian yang menjadi dasar pengembangan model prediksi depresi mahasiswa dalam penelitian ini. Ringkasan penelitian terdahulu tersebut disajikan pada tabel berikut.

Tabel 2.1 Tabel Penelitian Terdahulu

No	Judul Artikel & Penulis	Objek Penelitian	Metodologi	Temuan
1	Arafat, K. M. Y., Hossain, A., Ikfat, M., Amin, M. A., Tanvir, K., Gomes, D., & Rahman, M. (2026). *FRF-HHO: Early ovarian cancer prediction using explainable fuzzy random forest optimized by Harris Hawks algorithm*. Computers in Biology and Medicine, 192, 110536.	Prediksi kanker ovarium stadium awal menggunakan data klinis dan biomarker (349 pasien, 51 fitur)	Fuzzy Random Forest (FRF) yang dioptimasi dengan Harris Hawks Optimization (HHO), RFECV untuk seleksi fitur, SMOTE-Tomek untuk balancing, serta interpretasi menggunakan SHAP dan LIME	Model FRF-HHO mencapai akurasi 94.12%, recall 96.07%, dan F1-score 93.69%. Metode ini meningkatkan sensitivitas dan interpretabilitas model dibandingkan ensemble standar
2	Khan, A., Rahman, M. M., Ahmed, S., & Islam, M. R. (2023). *Predicting the depression in university students using stacking ensemble techniques over oversampling method*. Heliyon, 9(5), e15964.	Prediksi depresi pada mahasiswa universitas	Stacking ensemble dengan teknik oversampling untuk menangani ketidakseimbangan data	Teknik stacking dengan oversampling meningkatkan performa klasifikasi depresi dibandingkan model tunggal

3	Kessler, R. C., et al. (2016). *Machine learning-based prediction of illness course in major depression: The relevance of risk factors.*	Pasien dengan depresi mayor	Berbagai algoritma *machine learning*, analisis faktor risiko	Model mampu mengidentifikasi faktor risiko signifikan yang memengaruhi perjalanan penyakit depresi.
4	Akter, S., Liu, Z., Simoes, E. J., & Rao, P. (2025). *Using machine learning and electronic health record (EHR) data for the early prediction of Alzheimer's Disease and Related Dementias.*	Data rekam medis elektronik (EHR) pasien usia ≥40 tahun	GBT, LightGBM, *Random Forest*, XGBoost, Logistic Regression, AdaBoost; evaluasi menggunakan *AUC-ROC*	GBT memperoleh *AUC* tertinggi (0.809–0.833). Faktor risiko termasuk depresi dan kecemasan.
5	Panicker, T. F., Sarkar, D., Krishna, R., Mishra, R. K., Manjeshwar, S. K., & Sharma, A. (2026). *A machine learning-assisted prediction of potential biochar yield subjected to the physicochemical properties of biomass.*	Prediksi hasil *biochar* dari biomassa	*Random Forest Regression*, GBR, XGBR, SVR, dll.; analisis *partial dependence*	Model RF dan GBR mencapai $R^2 > 0.90$; suhu menjadi factor paling berpengaruh.
6	Xu, S., Zhou, Z., Sun, L., Nie, P., Hashimoto, S., & Jiang, W. (2026). *Random forest-based speed forecasting and carbon capture optimization for shipboard integrated energy systems.*	Sistem energi kapal dan prediksi kecepatan	*Random Forest* (100 decision trees), optimasi sistem energi terintegrasi	RF meningkatkan akurasi prediksi beban kecepatan dan menurunkan emisi karbon 12–15%.
7	Soladoye, A. A., Aderinto, N., Omodunbi, B. A., Esan, A. O., Adeyanju, I. A., & Olawade, D. B. (2025). *Enhancing Alzheimer's disease prediction using random forest: A novel framework combining backward feature elimination and ant colony optimization*. Current Research in Translational Medicine, 73, 103526.	Prediksi penyakit Alzheimer berbasis data klinis (2.149 data, 34 fitur).	Random Forest dikombinasikan dengan Backward Feature Elimination (BFE) dan Ant Colony Optimization (ACO), serta SMOTE untuk balancing data.	Kombinasi BFE-ACO menghasilkan akurasi 95% dan AUC 98%, lebih unggul dibanding metode konvensional serta lebih efisien secara komputasi.
8	Masud, G. H. A., Shanto, R. I., Sakin, I., & Kabir, M. R. (2025). *Effective depression detection and interpretation: Integrating machine learning, deep learning,*	Deteksi depresi pada mahasiswa menggunakan data sosial, demografis, dan teks.	Integrasi Machine Learning (Random Forest), Deep Learning, Transformer-based Language Models (BERT,	Random Forest mencapai akurasi 91.1% dan F1-score 91.6%. Model bahasa seperti RoBERTa menunjukkan

	language models, and explainable AI. Array, 25, 100375.		RoBERTa), serta Explainable AI (SHAP & LIME).	recall tinggi (98.6%) dengan interpretasi berbasis XAI.
9	Abebe, E., Mulugeta, A., Madakkattel, I., Stacey, D., & Hyppönen, E. (2026). *Uncovering plasma protein biomarkers linked to depression: A differential abundance analysis and Mendelian randomization using large-scale data*. Journal of Affective Disorders, 399, 121134.	Identifikasi biomarker protein plasma terkait depresi (48.378 partisipan UK Biobank, 2.920 protein).	Differential Protein Abundance Analysis (DPAA), analisis jaringan (PPI), Functional Enrichment, dan Mendelian Randomization (MR).	Ditemukan 22 protein signifikan terkait depresi dan terlibat dalam jalur inflamasi-imun. Bukti kausal terbatas namun berpotensi sebagai target terapi.
10	Chan, L. E., Casiraghi, E., Reese, J., Harmon, Q. E., Schaper, K., Hegde, H., Valentini, G., Schmitt, C., Motsinger-Reif, A., Hall, J. E., Mungall, C. J., Robinson, P. N., & Haendel, M. A. (2024). *Predicting nutrition and environmental factors associated with female reproductive disorders using a knowledge graph and random forests*. International Journal of Medical Informatics, 187, 105461.	Prediksi faktor nutrisi dan lingkungan terkait gangguan reproduksi wanita (9.765 responden).	Knowledge Graph untuk integrasi data ontologi dan Random Forest untuk prediksi hubungan faktor risiko.	Menghasilkan 8.535 prediksi hubungan signifikan/sugestif antara faktor lingkungan dan gangguan reproduksi; mendukung pembentukan hipotesis klinis.

Kesepuluh artikel pada tabel menunjukkan keterhubungan yang kuat dari perspektif permasalahan, metode, dan hasil. Dari sisi masalah, seluruh penelitian berfokus pada prediksi dan identifikasi faktor risiko dalam domain kesehatan dan sistem terapan, seperti prediksi kanker ovarium pada studi Arafat, prediksi depresi mahasiswa pada Khan dan Masud, serta deteksi Alzheimer pada Akter et dan Soladoye [1], [3], [7], [11], [12], . Selain itu, Abebe mengeksplorasi biomarker protein plasma terkait depresi, sementara Chan menganalisis faktor nutrisi dan lingkungan terhadap gangguan reproduksi wanita [10], [13]. Di luar ranah klinis langsung, Panicker dan Xu menunjukkan bahwa pendekatan machine learning juga efektif pada prediksi berbasis data fisik dan sistem energi [15], [16]. Dari sisi metode, mayoritas studi memanfaatkan Random Forest dan variannya, baik dalam

bentuk optimasi (FRF-HHO), kombinasi feature selection (BFE, ACO), ensemble (stacking), maupun integrasi dengan model lain seperti GBT, XGBoost, dan transformer (BERT, RoBERTa). Dari sisi hasil, seluruh penelitian menunjukkan peningkatan performa signifikan, baik dalam akurasi, AUC, recall, maupun efisiensi sistem, sekaligus menekankan pentingnya interpretabilitas melalui SHAP, LIME, atau analisis faktor risiko.

Berdasarkan sepuluh artikel tersebut, penelitian ini mengadopsi beberapa pendekatan utama yang relevan dengan pengembangan model prediksi depresi mahasiswa. Penggunaan algoritma *Random Forest* sebagai model utama merujuk pada performa tinggi yang ditunjukkan pada penelitian Arafat, Soladoye, serta Masud dalam tugas klasifikasi data kesehatan mental [1], [7], [12]. Selain itu, penelitian ini juga menerapkan metode *Synthetic Minority Oversampling Technique (SMOTE)* untuk menangani ketidakseimbangan data, yang diadaptasi dari pendekatan *oversampling* pada penelitian Khan dan Soladoye [3], [7]. Dalam proses seleksi fitur, penelitian ini memanfaatkan metode *Recursive Feature Elimination with Cross Validation (RFECV)* yang terinspirasi dari pendekatan optimasi fitur seperti *Backward Feature Elimination* dan *Ant Colony Optimization* yang digunakan oleh Arafat dan Soladoye [1], [7]. Tidak hanya berfokus pada performa model, penelitian ini juga mempertimbangkan aspek interpretabilitas melalui penerapan *Explainable Artificial Intelligence (XAI)* menggunakan metode *SHAP* dan *LIME*, sebagaimana diterapkan pada penelitian Arafat dan Masud [1], [12]. Dengan mengintegrasikan algoritma klasifikasi, teknik penyeimbangan data, seleksi fitur, serta interpretasi model dalam satu kerangka kerja, penelitian ini diharapkan mampu menghasilkan model prediksi depresi mahasiswa yang tidak hanya memiliki performa yang baik, tetapi juga lebih transparan, adaptif, dan mudah dipahami dalam implementasi praktis.

Adapun kebaruan penelitian ini dibandingkan penelitian terdahulu terletak pada integrasi pendekatan metodologis yang lebih terfokus pada konteks dan karakteristik data yang digunakan. Kebaruan penelitian ini berupa penerapan kombinasi metode yang meliputi algoritma *Random Forest* sebagai model

klasifikasi utama, penggunaan *Synthetic Minority Over-sampling Technique* (*SMOTE*) untuk mengatasi ketidakseimbangan kelas, seleksi fitur menggunakan *Recursive Feature Elimination with Cross Validation* (*RFECV*), serta interpretasi model menggunakan *SHapley Additive exPlanations* (*SHAP*) dan *Local Interpretable Model-agnostic Explanations* (*LIME*). Kombinasi metode tersebut digunakan untuk menghasilkan model prediksi yang tidak hanya memiliki performa klasifikasi yang baik, tetapi juga mampu memberikan penjelasan terhadap faktor-faktor yang memengaruhi hasil prediksi.

Pendekatan ini dipilih karena dalam pengembangan model prediksi, tidak hanya diperlukan tingkat *accuracy* yang tinggi, tetapi juga kemampuan model dalam melakukan generalisasi serta memberikan interpretasi yang jelas terhadap hasil prediksi. Jika Arafat dan Soladoye berfokus pada optimasi model melalui pendekatan *metaheuristic*, serta Masud menekankan integrasi *deep learning* dan *language model*, maka penelitian ini menghadirkan kombinasi metode yang disesuaikan secara spesifik dengan kebutuhan prediksi pada domain yang diteliti, dengan penekanan pada keseimbangan antara *accuracy*, *generalization*, dan *interpretability* [1], [7], [12]. Selain itu, berbeda dengan Abebe yang berorientasi pada analisis *biomarker* skala besar atau Chan yang memanfaatkan *knowledge graph*, penelitian ini lebih menitikberatkan pada konstruksi model prediktif yang efisien, transparan, dan aplikatif untuk implementasi praktis dalam sistem deteksi dini berbasis data [10], [13]. Dengan demikian, kontribusi utama penelitian ini tidak hanya terletak pada performa model, tetapi juga pada integrasi metodologis yang lebih adaptif dan kontekstual dibandingkan penelitian-penelitian sebelumnya.

2.2 Teori yang berkaitan

2.2.1 Depresi

Depresi merupakan gangguan suasana hati (*mood disorder*) yang ditandai dengan perasaan sedih berkepanjangan, kehilangan minat terhadap aktivitas sehari-hari, gangguan tidur, perubahan nafsu makan, serta kesulitan konsentrasi [13]. Pada populasi mahasiswa, depresi sering berkaitan dengan

tekanan akademik, tuntutan sosial, dan faktor psikososial lainnya. Kondisi ini dapat berdampak pada penurunan performa akademik serta kualitas hidup secara keseluruhan apabila tidak terdeteksi sejak dini. Urgensi deteksi dini depresi juga diperkuat oleh berbagai penelitian terdahulu. Penelitian oleh Khan menunjukkan bahwa depresi pada mahasiswa dapat diprediksi menggunakan pendekatan *machine learning* berbasis data, sehingga proses identifikasi dapat dilakukan secara lebih cepat dan objektif [3].

Sementara itu, studi Masud membuktikan bahwa integrasi model pembelajaran mesin dan teknik interpretasi mampu meningkatkan akurasi deteksi depresi mahasiswa hingga di atas 90% [12]. Temuan-temuan tersebut menunjukkan bahwa depresi tidak hanya dapat dianalisis dari perspektif klinis, tetapi juga melalui pendekatan komputasional berbasis data. Dengan demikian, konsep depresi dalam penelitian ini tidak hanya dipahami sebagai fenomena psikologis, tetapi juga sebagai permasalahan yang dapat dimodelkan secara prediktif menggunakan teknik analisis data [14].

2.2.2 Machine Learning

Machine Learning merupakan cabang dari kecerdasan buatan yang memungkinkan sistem untuk mempelajari pola dari data dan menghasilkan prediksi. Dalam konteks kesehatan mental, pendekatan ini termasuk dalam kategori *supervised learning*, di mana model dilatih menggunakan data yang telah memiliki label, seperti status depresi.

Berbagai penelitian terdahulu menunjukkan bahwa *machine learning* memiliki performa yang baik dalam prediksi kondisi kesehatan. Mengidentifikasi bahwa algoritma pembelajaran mesin mampu memetakan faktor risiko yang mempengaruhi perjalanan depresi mayor. Akter juga menunjukkan bahwa model berbasis pembelajaran mesin seperti Gradient Boosting dan Random Forest menghasilkan nilai AUC yang tinggi dalam prediksi gangguan kognitif menggunakan data rekam medis elektronik [11].

Hal ini menunjukkan bahwa pendekatan berbasis ML efektif dalam menangani data kompleks dengan banyak variabel.

Selain itu, penelitian Khan membandingkan beberapa teknik klasifikasi dan menemukan bahwa metode berbasis *ensemble* memberikan performa yang lebih baik dibanding model tunggal [3]. Temuan ini memperkuat bahwa dalam konteks prediksi depresi, pemilihan algoritma yang tepat sangat berpengaruh terhadap hasil model. Dengan demikian, penggunaan *machine learning* dalam penelitian ini didasarkan pada bukti empiris bahwa metode tersebut mampu mengidentifikasi pola kompleks dalam data kesehatan mental dan menghasilkan prediksi yang akurat.

2.2.3 Random Forest

Random Forest merupakan salah satu algoritma *ensemble learning* yang membangun sejumlah pohon keputusan (*decision tree*) dan menggabungkan hasilnya melalui mekanisme *majority voting* [15]. Pendekatan ini bertujuan untuk meningkatkan akurasi serta mengurangi risiko *overfitting* yang sering terjadi pada model tunggal. Berbagai penelitian terdahulu menunjukkan efektivitas algoritma ini dalam berbagai domain kesehatan. Soladoye melaporkan bahwa *Random Forest* yang dikombinasikan dengan teknik seleksi fitur mampu mencapai akurasi hingga 95% dan AUC sebesar 98% dalam prediksi Alzheimer [7]. Masud juga menunjukkan bahwa *Random Forest* memperoleh performa tinggi dalam deteksi depresi mahasiswa, dengan tingkat akurasi di atas 90% [12].

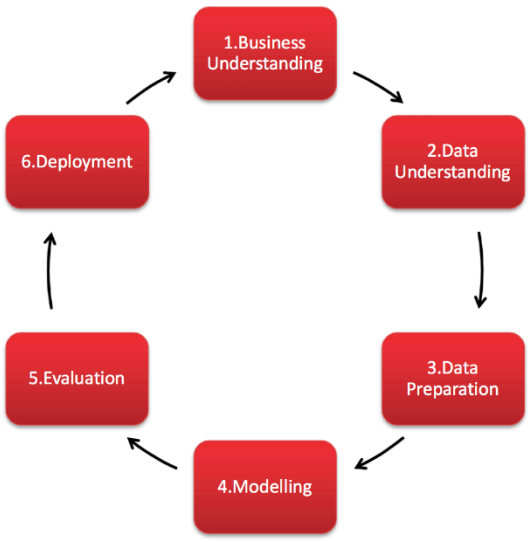
Sementara itu, Arafat membuktikan bahwa pengembangan model berbasis Random Forest yang dioptimasi dapat meningkatkan sensitivitas dalam deteksi dini penyakit [1]. Meskipun beberapa penelitian menggunakan optimasi metaheuristik atau integrasi dengan model lain, penelitian ini memilih menggunakan *Random Forest* sebagai algoritma utama dengan pendekatan yang lebih terfokus pada keseimbangan antara *accuracy*, *generalization*, dan *interpretability*. Pemilihan algoritma ini dalam penelitian bertujuan untuk membangun model prediksi depresi mahasiswa yang stabil,

akurat, dan mampu memberikan informasi mengenai variabel yang paling berpengaruh terhadap status depresi. Random Forest dipilih karena memiliki performa yang stabil pada data tabular, mampu menangani hubungan non-linear antar fitur, tahan terhadap overfitting melalui mekanisme bagging, mampu menangani kombinasi data numerik dan kategorikal, serta menyediakan feature importance yang penting untuk interpretasi faktor risiko depresi.

2.3 Framework/Algoritma yang digunakan

2.3.1 CRISP-DM

Dalam penelitian ini, framework yang digunakan adalah *Cross Industry Standard Process for Data Mining* (CRISP-DM), sedangkan algoritma yang diterapkan pada tahap pemodelan adalah *Random Forest*. Penggunaan CRISP-DM bertujuan untuk memastikan proses penelitian berjalan secara sistematis dan terstruktur, sementara *Random Forest* digunakan sebagai metode klasifikasi untuk memprediksi depresi mahasiswa.



Gambar 2.1 Tahapan CRISP-DM

1. Business Understanding

Tahap *Business Understanding* merupakan tahap awal yang bertujuan untuk memahami permasalahan penelitian secara

komprehensif. Dalam konteks penelitian ini, permasalahan yang diangkat adalah tingginya risiko depresi pada mahasiswa yang sering kali tidak terdeteksi secara dini. Depresi dapat berdampak pada penurunan performa akademik, gangguan sosial, serta risiko kesehatan mental yang lebih serius. Oleh karena itu, diperlukan suatu model prediksi berbasis data yang mampu mengidentifikasi mahasiswa yang berpotensi mengalami depresi. Tujuan utama penelitian ini adalah membangun model klasifikasi yang akurat menggunakan algoritma *Random Forest* untuk memprediksi status depresi berdasarkan variabel-variabel dalam dataset yang digunakan.

2. Data Understanding

Tahap *Data Understanding* dilakukan untuk memahami karakteristik dataset yang digunakan dalam penelitian. Pada tahap ini dilakukan identifikasi jumlah data, jumlah variabel, tipe data (numerik maupun kategorikal), serta distribusi kelas pada variabel target. Selain itu, dilakukan analisis awal terhadap kemungkinan adanya nilai kosong (*missing values*), data duplikat, maupun ketidakseimbangan kelas antara mahasiswa yang mengalami depresi dan tidak depresi. Pemahaman terhadap struktur dan kualitas data sangat penting karena akan memengaruhi proses pemodelan serta hasil evaluasi model.

3. Data Preparation

Tahap *Data Preparation* merupakan proses penyiapan data sebelum dilakukan pemodelan. Pada tahap ini dilakukan pembersihan data (*data cleaning*) untuk mengatasi nilai kosong, inkonsistensi data, maupun data yang tidak relevan. Variabel kategorikal diubah menjadi bentuk numerik menggunakan teknik *encoding* agar dapat diproses oleh algoritma *machine learning*. Apabila ditemukan ketidakseimbangan kelas, dilakukan penyesuaian agar model tidak bias terhadap kelas mayoritas. Selanjutnya, dataset dibagi menjadi data latih (*training data*) dan data uji (*testing data*) dengan proporsi tertentu untuk memastikan evaluasi model dilakukan secara objektif. Tahap ini bertujuan

menghasilkan dataset yang bersih dan siap digunakan dalam proses pelatihan model.

4. Modeling

Tahap *Modeling* merupakan tahap pembangunan model menggunakan algoritma *Random Forest*. Algoritma ini bekerja dengan membangun sejumlah pohon keputusan (*decision tree*) berdasarkan sampel data yang dipilih secara acak melalui metode *bootstrap sampling*. Pada setiap percabangan pohon, fitur dipilih secara acak untuk mengurangi korelasi antar pohon dan meningkatkan generalisasi model. Hasil prediksi dari seluruh pohon kemudian digabungkan menggunakan metode *majority voting* untuk menentukan kelas akhir. Dalam tahap ini juga dilakukan penyesuaian parameter (*hyperparameter tuning*) seperti jumlah pohon (*n_estimators*) dan kedalaman maksimum pohon (*max_depth*) untuk memperoleh performa model yang optimal.

5. Evaluation

Tahap Evaluation bertujuan untuk menilai kinerja model yang telah dibangun. Model diuji menggunakan data uji (testing data) yang sebelumnya tidak digunakan dalam proses pelatihan. Evaluasi dilakukan menggunakan beberapa metrik klasifikasi, seperti accuracy, precision, recall, F1-score, dan AUC (Area Under the Curve). Penggunaan berbagai metrik ini bertujuan untuk memberikan gambaran yang lebih komprehensif mengenai kemampuan model dalam mengklasifikasikan mahasiswa yang mengalami depresi dan tidak depresi. Tahap evaluasi ini memastikan bahwa model memiliki performa yang baik dan tidak mengalami overfitting.

6. Deployment

Tahap *Deployment* merupakan tahap akhir dalam CRISP-DM, di mana hasil model diinterpretasikan dan disusun dalam bentuk laporan penelitian. Dalam konteks penelitian ini, tahap *deployment* tidak berupa

implementasi sistem secara langsung, melainkan penyajian hasil analisis dan interpretasi model. Informasi mengenai variabel yang paling berpengaruh terhadap prediksi depresi dianalisis melalui fitur *feature importance* pada *Random Forest*. Hasil penelitian ini diharapkan dapat menjadi dasar pengembangan sistem deteksi dini depresi mahasiswa berbasis *machine learning* di masa mendatang.

Tabel 2.2 Perbandingan Framework

Aspek	CRISP-DM	KDD	SEMMA
Fokus Utama	Proses lengkap proyek *data mining* dari bisnis hingga implementasi	Proses penemuan pengetahuan dari data	Proses teknis analisis data berbasis pemodelan
Orientasi	Berorientasi pada kebutuhan bisnis/masalah	Berorientasi pada penemuan pola dalam data	Berorientasi pada teknik analisis dan pemodelan
Tahapan Utama	1. *Business Understanding* 2. *Data Understanding* 3. *Data Preparation* 4. *Modeling* 5. *Evaluation* 6. *Deployment*	1. Selection 2. Preprocessing 3. Transformation 4. Data Mining 5. Interpretation / Evaluation	1. Sample 2. Explore 3. Modify 4. Model 5. Assess
Cakupan Proses	Lengkap dan komprehensif	Fokus pada proses teknis pengolahan data	Lebih teknis dan spesifik untuk

			perangkat lunak tertentu
Sifat Proses	Iteratif dan fleksibel	Linear namun dapat berulang	Iteratif namun berfokus pada analisis statistik
Keterlibatan Aspek Bisnis	Sangat kuat	Terbatas	Tidak secara eksplisit menekankan aspek bisnis
Penggunaan Umum	Penelitian akademik dan industri lintas bidang	Penelitian berbasis data dan akademik	Banyak digunakan pada software SAS
Kesesuaian untuk Penelitian	Sangat sesuai karena mencakup pemahaman masalah hingga evaluasi model	Cukup sesuai tetapi kurang menekankan aspek implementasi	Kurang sesuai karena lebih berorientasi teknis dan software tertentu

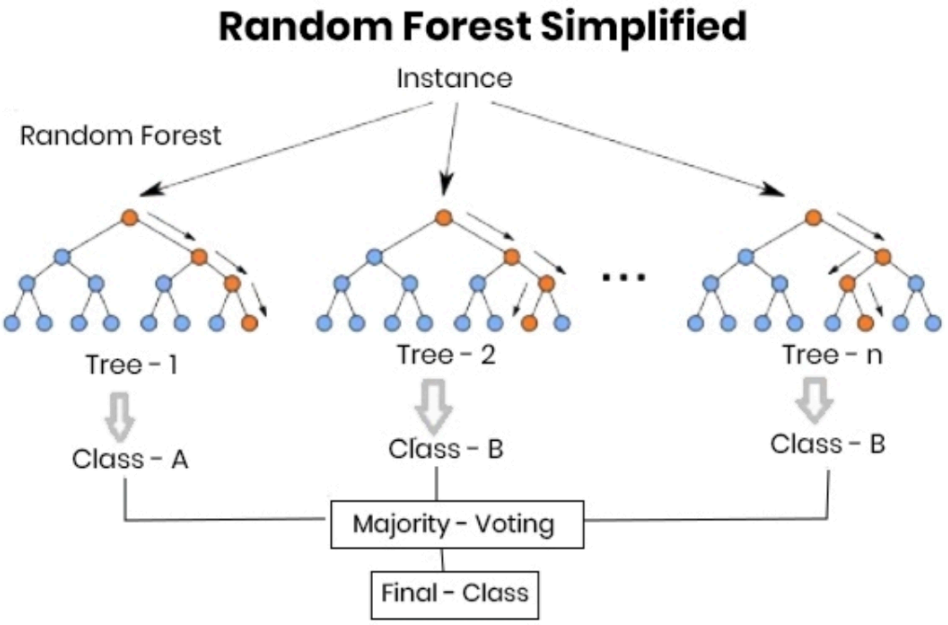
Berdasarkan Tabel 2.2 Perbandingan Framework, dapat dilihat bahwa ketiga metodologi tersebut memiliki tujuan yang sama, yaitu mendukung proses pengolahan dan analisis data untuk menghasilkan pengetahuan atau model prediksi. Namun, terdapat perbedaan mendasar pada orientasi dan cakupan proses masing-masing metodologi. KDD (*Knowledge Discovery in Databases*) lebih berfokus pada tahapan teknis dalam menemukan pola dari data, dimulai dari seleksi hingga interpretasi hasil. SEMMA lebih menekankan pada proses analisis statistik dan pemodelan data, serta umumnya digunakan dalam lingkungan perangkat lunak tertentu seperti SAS. Sementara itu, CRISP-DM memiliki cakupan yang lebih komprehensif karena

tidak hanya mencakup aspek teknis pengolahan data, tetapi juga dimulai dari pemahaman permasalahan bisnis hingga tahap implementasi (*deployment*).

Penelitian ini memilih CRISP-DM sebagai framework karena metodologi ini memberikan alur kerja yang sistematis, terstruktur, dan sesuai dengan kebutuhan penelitian prediksi depresi mahasiswa. CRISP-DM memungkinkan peneliti untuk memahami permasalahan secara mendalam pada tahap *Business Understanding*, memastikan kualitas data pada tahap *Data Preparation*, serta melakukan evaluasi model secara komprehensif sebelum hasil penelitian disajikan. Selain itu, sifatnya yang iteratif memungkinkan perbaikan pada setiap tahap apabila ditemukan kekurangan selama proses pemodelan. Dengan demikian, CRISP-DM dinilai paling sesuai untuk mendukung pengembangan model prediksi depresi menggunakan algoritma *Random Forest* dalam penelitian ini.

2.3.2 Random Forest

Random Forest merupakan salah satu algoritma dalam kategori *ensemble learning* yang digunakan untuk permasalahan klasifikasi maupun regresi. Algoritma ini dikembangkan oleh Leo Breiman pada tahun 2001 sebagai pengembangan dari metode *Decision Tree*. Konsep utama dari *Random Forest* adalah menggabungkan banyak pohon keputusan (*decision trees*) yang dibangun dari sampel data dan subset fitur yang dipilih secara acak, kemudian mengombinasikan hasilnya untuk memperoleh prediksi akhir yang lebih akurat dan stabil.



Gambar 2.2 Algoritma Random Forest

Secara umum, kelemahan dari *Decision Tree* tunggal adalah rentan terhadap *overfitting*, yaitu kondisi ketika model terlalu menyesuaikan diri dengan data latih sehingga performanya menurun pada data baru. Untuk mengatasi permasalahan tersebut, *Random Forest* menggunakan pendekatan agregasi yang dikenal sebagai *bagging* (*bootstrap aggregating*). Pada metode ini, beberapa sampel data diambil secara acak dengan pengembalian (*sampling with replacement*) untuk membangun sejumlah pohon keputusan yang berbeda.

Misalkan tersedia dataset pelatihan:

$$D=\{(x1,y1),(x2,y2),...,(xn,yn)\}$$

Rumus 2.1

di mana:

- x_i adalah vektor fitur ke-i
- y_i adalah label kelas ke-i

- $n$adalah jumlah data

1. Proses Bootstrap Sampling

   Pada tahap pertama, dilakukan pembentukan *B* subset data latih secara acak dari dataset asli menggunakan metode *bootstrap*. Setiap subset digunakan untuk membangun satu pohon keputusan. Dengan demikian akan terbentuk sejumlah *B* pohon keputusan:

   T1(x),T2(x),T3(x),...,TB(x)

   Rumus 2.2

   di mana $T_b(x)$adalah prediksi dari pohon ke-b.

2. Pemilihan Fitur Secara Acak

   Pada setiap node dalam pohon keputusan, hanya sebagian fitur yang dipilih secara acak sebanyak $m$dari total $p$fitur yang tersedia (dengan $m < p$). Hal ini bertujuan untuk mengurangi korelasi antar pohon dan meningkatkan keberagaman model.

3. Prediksi pada Random Forest (Klasifikasi)

   Untuk kasus klasifikasi, prediksi akhir diperoleh melalui metode *majority voting*, yaitu kelas yang paling banyak dipilih oleh seluruh pohon keputusan.

   Secara matematis, prediksi akhir $\hat{y}$untuk suatu data $x$dapat dituliskan sebagai:

   $$\hat{y} = mode\{T_1(x), T_2(x),..., T_B(x)\}$$

   $Rumus\ 2.3$

   Artinya, kelas yang memiliki frekuensi terbanyak dari seluruh prediksi pohon menjadi hasil akhir klasifikasi.

4. Prediksi pada Random Forest (Regresi)

Untuk kasus regresi, prediksi akhir diperoleh dari rata-rata hasil prediksi seluruh pohon:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

Rumus 2.4

Namun, dalam penelitian ini digunakan pendekatan klasifikasi karena tujuan penelitian adalah memprediksi kategori depresi (ya/tidak).

5. Fungsi Impurity pada Decision Tree

Pada setiap pohon dalam *Random Forest*, pemilihan percabangan node biasanya menggunakan ukuran impurity seperti *Gini Index*. Rumus *Gini Index* adalah:

di mana:

$$Gini = 1 - \sum_{i=1}^{k} p_i^2$$

Rumus 2.5

- p_i adalah proporsi kelas ke-i dalam suatu node
- k adalah jumlah kelas

Semakin kecil nilai Gini, maka node tersebut semakin homogen.

6. Keunggulan Random Forest

Beberapa keunggulan *Random Forest* dalam penelitian ini adalah:

1. Mengurangi risiko *overfitting* karena menggunakan banyak pohon.
2. Memiliki akurasi yang tinggi dalam permasalahan klasifikasi biner.
3. Mampu menangani data dengan jumlah fitur yang banyak.

4. Dapat menghitung tingkat kepentingan variabel (*feature importance*).
5. Tidak sensitif terhadap skala data.

Dalam penelitian ini, *Random Forest* digunakan untuk mengklasifikasikan mahasiswa ke dalam dua kategori, yaitu mengalami depresi atau tidak mengalami depresi. Model dibangun menggunakan data latih dan dievaluasi menggunakan metrik seperti *accuracy*, *precision*, *recall*, *F1-score*, dan *AUC* untuk memastikan performa yang optimal.

2.4 Tools/software yang digunakan

Dalam penelitian ini, proses pengolahan data, pembangunan model, serta evaluasi dilakukan menggunakan perangkat lunak berbasis *open-source*. Pemilihan tools dilakukan dengan mempertimbangkan kemudahan implementasi algoritma *machine learning*, ketersediaan pustaka pendukung yang lengkap, serta fleksibilitas dalam eksplorasi dan visualisasi data. Adapun tools utama yang digunakan dalam penelitian ini adalah bahasa pemrograman Python dan lingkungan kerja Jupyter Notebook.

2.4.1 Bahasa Pemrograman Python

Python merupakan bahasa pemrograman tingkat tinggi (*high-level programming language*) yang banyak digunakan dalam bidang *data science*, *machine learning*, dan kecerdasan buatan. Python dipilih dalam penelitian ini karena memiliki sintaks yang sederhana, mudah dipahami, serta didukung oleh komunitas pengembang yang besar. Selain itu, Python menyediakan berbagai pustaka (*libraries*) yang mendukung proses analisis data dan pembangunan model prediksi.

Gambar 2.3 Bahasa Pemrograman *Python*

Beberapa pustaka utama yang digunakan dalam penelitian ini antara lain:

1. Pandas, digunakan untuk proses manipulasi dan analisis data, seperti membaca dataset, membersihkan data, serta melakukan transformasi variabel.
2. NumPy, digunakan untuk operasi numerik dan pengolahan array multidimensi.
3. Scikit-learn, digunakan untuk implementasi algoritma *Random Forest*, pembagian data latih dan data uji, serta evaluasi model menggunakan metrik seperti *accuracy*, *precision*, *recall*, *F1-score*, dan *AUC*.
4. Matplotlib dan Seaborn, digunakan untuk visualisasi data dan hasil evaluasi model.

Penggunaan Python dalam penelitian ini memungkinkan proses implementasi framework CRISP-DM dilakukan secara sistematis, mulai dari tahap *data preparation* hingga *evaluation*. Selain itu, Python juga mendukung analisis *feature importance* yang menjadi salah satu keunggulan algoritma *Random Forest* dalam mengidentifikasi variabel yang paling berpengaruh terhadap prediksi depresi.

2.4.2 Jupyter Notebook

Jupyter Notebook merupakan lingkungan pengembangan interaktif (*interactive development environment*) yang memungkinkan peneliti untuk menulis dan menjalankan kode program secara bertahap dalam bentuk sel (*cell-based execution*). Jupyter Notebook dipilih dalam penelitian ini karena memudahkan proses eksplorasi data, dokumentasi kode, serta penyajian hasil analisis dalam satu dokumen yang terintegrasi.

Gambar 2.4 Jupyter Notebook

Keunggulan *Jupyter Notebook* dalam penelitian ini antara lain:

1. Mendukung eksekusi kode secara bertahap sehingga memudahkan proses *debugging* dan eksperimen model.
2. Memungkinkan penggabungan antara kode program, teks penjelasan, dan visualisasi dalam satu dokumen.
3. Memudahkan proses eksplorasi data (*exploratory* data *analysis*).
4. Mendukung berbagai pustaka *Python* yang digunakan dalam pembangunan model *Random Forest*.

Dengan menggunakan *Jupyter Notebook*, seluruh proses penelitian mulai dari pemuatan dataset, pembersihan data, pelatihan model, hingga evaluasi hasil dapat terdokumentasi secara sistematis dan transparan. Hal ini mendukung prinsip reprodusibilitas penelitian (*research reproducibility*), di mana proses analisis dapat diulang dan diverifikasi kembali apabila diperlukan.